INFORMATION RETRIEVAL PERFORMANCE ENHANCEMENT USING THE AVERAGE STANDARD ESTIMATOR AND THE MULTI-CRITERIA DECISION WEIGHTED SET OF PERFORMANCE MEASURES

by

TAREQ Z. AHRAM
B.Sc., Industrial Eng., University of Jordan, Jordan, 2002
M.Sc., Industrial Engineering and Management, University of Jordan, Jordan, 2004
M.Sc., Industrial Engineering, University Central Florida, USA, 2007

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Industrial Engineering and Management Systems
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2008

Major Professor: Dr. Pamela McCauley-Bush

ABSTRACT

Information retrieval is much more challenging than traditional small document collection retrieval. The main difference is the importance of correlations between related concepts in complex data structures. These structures have been studied by several information retrieval systems. This research began by performing a comprehensive review and comparison of several techniques of matrix dimensionality estimation and their respective effects on enhancing retrieval performance using singular value decomposition and latent semantic analysis. Two novel techniques have been introduced in this research to enhance intrinsic dimensionality estimation, the Multi-criteria Decision Weighted model to estimate matrix intrinsic dimensionality for large document collections and the Average Standard Estimator (ASE) for estimating data intrinsic dimensionality based on the singular value decomposition (SVD). ASE estimates the level of significance for singular values resulting from the singular value decomposition. ASE assumes that those variables with deep relations have sufficient correlation and that only those relationships with high singular values are significant and should be maintained. Experimental results over all possible dimensions indicated that ASE improved matrix intrinsic dimensionality estimation by including the effect of both singular values magnitude of decrease and random noise distracters. Analysis based on selected performance measures indicates that for each document collection there is a region of lower dimensionalities associated with improved retrieval performance. However, there was clear disagreement between the various performance measures on the model associated with best performance. The introduction of the multi-weighted model and Analytical Hierarchy Processing (AHP) analysis helped in ranking dimensionality estimation techniques and facilitates satisfying

overall model goals by leveraging contradicting constrains and satisfying information retrieval priorities. ASE provided the best estimate for MEDLINE intrinsic dimensionality among all other dimensionality estimation techniques, and further, ASE improved precision and relative relevance by 10.2% and 7.4% respectively. AHP analysis indicates that ASE and the weighted model ranked the best among other methods with 30.3% and 20.3% in satisfying overall model goals in MEDLINE and 22.6% and 25.1% for CRANFIELD. The weighted model improved MEDLINE relative relevance by 4.4%, while the scree plot, weighted model, and ASE provided better estimation of data intrinsic dimensionality for CRANFIELD collection than Kaiser-Guttman and Percentage of variance. ASE dimensionality estimation technique provided a better estimation of CISI intrinsic dimensionality than all other tested methods since all methods except ASE tend to underestimate CISI document collection intrinsic dimensionality. ASE improved CISI average relative relevance and average search length by 28.4% and 22.0% respectively. This research provided evidence supporting a system using a weighted multi-criteria performance evaluation technique resulting in better overall performance than a single criteria ranking model. Thus, the weighted multi-criteria model with dimensionality reduction provides a more efficient implementation for information retrieval than using a full rank model.

I dedicate my dissertation work to my beloved family, and specifically, a special feeling of gratitude to my loving parents for their encouragement and support. My sisters and brothers have never left my side and are very special.

I also dedicate this dissertation to my friends for their support throughout my quest for knowledge.

ACKNOWLEDGMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Dr. Pamela McCauley-Bush, my committee chair, for her countless hours of advising, reading, encouraging and, most of all, her cheerful personality and patience throughout the entire process. I would also like to thank Drs. Ahmad Elshennawy, Yan Wang, Dima Nazzal, and Nabeel Yousef for agreeing to serve on my committee and for the invaluable feedback and comments. I would further like to thank Drs. Jason Dowling, Miles Efron and Dimitris Zeimpekis for their support and research assistance.

Finally, I would like to thank the instructors, mentors and professors in the department of Industrial Engineering and Management Systems for assisting me with this research with a special thanks to Dr. Waldemar Karwowski for his encouragement and support. My colleagues', Joylene Ware and Christopher Lee, excitement and willingness to provide feedback and support made the completion of this research an enjoyable experience.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

CHAPTER ONE: INTRODUCTION

1.1 Background

Internet users continue to rely on search engines as the primary way for finding information on the web. Results generated from search engines satisfy all kinds of information needs, ranging from scientific research to locating a place of interest to compare products and services. In the current web search engines, the process of identifying relevant documents usually involves matching queries with the keyword found in document databases located in the system data stores. That is, for a returned result to be considered relevant to search queries it has to contain some or all of the query keywords. This approach in searching for information has been successful in satisfying most of the user needs. However, there are some queries for which basic keyword matching will not be sufficient.

The purpose of Information Retrieval (IR) systems and search engines is to help people locate relevant information when a request it is made. An ambiguous query might be encountered because it is associated with more than one interpretation and each interpretation might be related to a different field of knowledge. Consequently, web pages that have different domains of knowledge, but all shares similar keywords, will be presented to users leaving them with the burden of filtering their search results. Resolving such problem has been for a long time has been the primary focus of many fields. It was estimated that the World Wide Web involved at least 350 million documents of different types and formats to nearly 800 million Web pages (Nielsen/Net Rating, 2000) (Lawrence & Giles, 1999). These documents were growing at the rate of 20 million per month, while

internet traffic volume continues to double about every 100 days (Computer Industry Almanac Inc., 2002).

Although many of the traditional IR techniques are useful, information retrieval from the web involves some issues. The estimated size of indexed web collections was at least 11.5 billion pages by the end of January 2005 (Gulli et al., 2005). To get a better understanding about the process of searching on the web, it is vital to have a clear idea about the size of the document collections involved in the search process. To facilitate comparison between various search engine providers, the document collection sizes in different search engines are displayed in Figure 1.

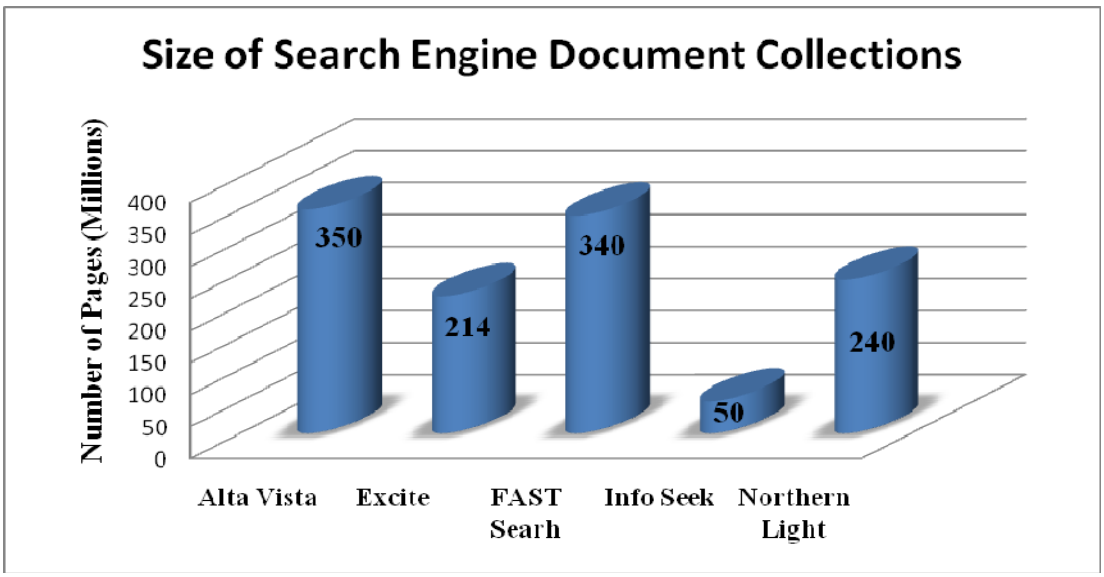


Figure 1: Indexed Search Engines Document Collections (millions of pages)

(Source: Jansen, B. J. 2000)

Figure 1 clearly indicates that the difference in the size of indexed documents has a great impact on web searching. Because of the huge size and dynamic growth of these document collections, users can easily be distracted with various returned results (Xu, 1999).

Research on Information Retrieval systems is based on small controlled collections of scientific data repositories on a particular topic (Brin and Page, 1998). The nature of the World Wide Web is also different from traditional Information Retrieval systems. Web IR includes digital pictures; video and audio data in addition to text from different languages which is found on frequently duplicated web pages (Huang, 2000). Additionally, web search engines and information retrieval systems are frequently affected by external factors which try to manipulate search engine responses (Brin and Page, 1998). Further problem is the number of queries which a search engine might have to handle, in the case of Google search engine this is thousands of queries per second (Brin and Page, 1998). In a recent research conducted by Nielsen/Ratings (2006), Google's searches increased from nearly 2.1 billion in March 2005 to 2.9 billion in March 2006 this is shown in Table 1. Currently several search engines add popularity to link analysis methods and consequently the application of link usage to collect information to determine relevance and popularity of web pages – thus the more often web pages are entered by users, the higher their relevance (Liddy, 2001).

Table 1: Search Engines Growth for Top Ranking Internet Search Providers

(*Source: Nielsen/NetRatings Mega View Search, April 2006*)

Provider	Mar-05 Searches	Mar-06 Searches	Year-over-Year Percent Change
Google Search	2,057,897,000	2,900,375,000	+41%
Yahoo! Search	907,751,000	1,330,183,000	+47%
MSN Search	592,153,000	643,803,000	+9%

There are three main techniques which have been proposed for IR (Salton, 1989): the Boolean model; which consists of separating keywords with Boolean expressions such as "AND" and "OR", a Probabilistic model based on relevance of the documents in the Vector Space Model (VSM). The Boolean logic has been used for early commercial systems. VSM, which will be discussed next, is more precise and is simpler and easier to implement (Baeza-Yates and Ribeiro-Neto, 1999). Latent Semantic Indexing (LSI) is an extension to the VSM. LSI is an attempt to match the meaning of a document to user query by locating documents with similar properties closer together in a vector space. Past performance results, which are presented in Chapter 2, shows that LSI method is a better indicator of meaning in a document than individual terms. LSI is performed by using a numerical computation technique called Singular Value Decomposition (SVD).

1.2 Vector Space Modeling (VSM)

Salton's Vector Space Model (VSM) treats documents as vectors in a dimensional space with inter-document similarity represented by their corresponding vector cosine (Salton et al., 1983). Documents that are about similar topics lie near each other. Thus information retrieval is concerned with navigating this vector space; while attempting to locate regions of the vector space that contain documents relevant to specific information needs.

Improvements on Salton's model, known as the generalized vector space model (GVSM) (Wong, et al., 1987) have suggested that alternatives to this vector space may be beneficial. Due to the non-orthogonality and interdependence of natural language terms, such model of the observed term space relations may improve retrieval.

Latent Semantic Indexing (LSI) introduce the vector space orthogonal projection of its P-dimensional document vectors onto a k-dimensional subspace, where in LSI ($k < p$).

Dimensionality reduction provides a systematic representation of term-document associations, similar objects are arranged by eliminating observed data over specification error (Deerwester et al., 1990). My research is concerned with the parameterization of *k*, the number of dimensions retained during the implementation of LSI orthogonal projection while satisfying a set of weighted performance measures. This research is aimed at discovering a better and more effective means for selecting *k* in unsupervised environments while maintaining a reasonable query response time for information retrieval systems.

This research will try to answer the following question: Can we get better search results in terms of relevance and precision, while reducing search response time through the use of selected dimensionality reduction parameter in the truncated singular value decomposition?

LSI reflects terms and documents in an orthogonal subspace of the term-document matrix *A* by means of the singular value decomposition (SVD). Matrix dimensionality reduction calculates what is called "singular values" of A, which are the positive square roots of the eigenvalues of *A'A*.

1.3 Information Retrieval Aboutness and Relevance

An information space is the set of concepts and relations between them held by a computer system (Newby, 2001). In the field of cognitive science, the philosophical status of concepts is a matter of ongoing debate (Laurence et al., 1999) (Rosch, 1999) (Quine, 1999). Measuring Aboutness and Relevance in information space in not typically open to observations or direct notice, Hutchins (1978) introduced the concept of "*Aboutness*". Without assurance of *Aboutness,* the *Relevance* of a document to a query is hard to check. Relevance between documents and queries is closely tied to a third representation in IR

problem which is *Similarity*. Documents that are relevant to a query are in some way similar to it, and relevant documents are similar to each other. Aboutness, Relevance, and Similarity are all important to IR technologies. According to Gardenfors (2000), concepts contain variables that measure the properties of objects. An information space could be described as the set of variables observed by a system and the system means of associating them. Thus mass, volume, and density might be concepts in an information space related to physical measurements. On the other hand radiation and convection might be important concepts in the information space of an IR system related to energy transfer. Accordingly, dimensions provide the structure of the space and define the form that informs common notions of similarity and distance. As Gardenfors writes, "*dimensions form the framework used to assign properties to objects and to specify relations among them. The coordinates of a point within a conceptual space represent particular instances of each dimension..."* (Gardenfors, 2000)

The assumption of term independence is a major problem in VSM. In Salton's model, documents contained in the information space spanned by the system's indexing terms, and similarity is defined by the vector cosine. Thus if car and automobile are both present in the indexing vocabulary, systems based on the standard vector space model will fail to retrieve documents indexed on automobiles for queries about cars. To see why this is the case, consider the similarity function of the VSM given an $n \times p$ document-term matrix A and a p-dimensional query vector q, VSM similarity function is given in Equation 1.3.1 .

(1.3.1) $$s = qA'$$

In Equation 1.3.1, s is the n-vector of similarity values. Under the standard VSM, dimensions of term space are assumed to be orthogonal; the model assumes that terms are

statistically independent. Assumption of term independence may be covered by re-writing Equation 1.3.1 as shown in Equation 1.3.2 (Jiang et al., 2000).

(1.3.2) $$s = qI_p A^{'}$$

In Equation 1.3.2, the identity matrix I_p covers independence among the indexing variables. Wong et al. (1987), suggest that term correlation information should be reflected in the model. Wong extended Salton's vector space theory and proposed the generalized vector space model (GVSM). In Equation 1.3.3, *R* is the p x p term correlation matrix for A. Thus; according to the GVSM, if "home" and "house" tend to co-occur in an information space, an IR system will reflect their relationship in matrix *R*.

(1.3.3) $$s_{GVSM} = qRA^{'}$$

This sample correlation matrix *R* provides a model of the relationships between indexing terms. GVSM attempts to improve Salton's model by allowing information space to include inter-term correlation. That is, by replacing the identity matrix of Equation 1.3.2 with the correlation matrix *R* in Equation 1.3.3, the GVSM minimize the error introduced to the Salton's VSM by assuming term independence. Overall, Salton's VSM deviates from reality by assuming simplicity when VSM suggested statistical independence among terms. Generalized Vector Space Model (GVSM) removes error from Salton's Vector Space Model (VSM) theory by including the observed term correlations. Latent Semantic Indexing (LSI) removes error from the GVSM through a model based on the observed sample of the population correlation matrix. In LSI, we have the similarity function:

(1.3.4) $$s_{LSI} = qR_k A^{'}$$

In Equation 1.3.4, R_k is the rank- k approximation of R according to the least-squares method, where $k \leq fullrank(A)$. Equation 1.3.4 adds to the traditional VSM a reduced linear model of the correlational arrangement of the indexing (terms) found in A. Selecting the best value of k that returns in a better query results have been till these days a problem of statistical model building that was not covered extensively in most IR research with Matrix Decomposition and LSI.

While Wong GVSM adds to Salton's model by including a model of term association based on the sample correlation matrix. LSI's improvement over Salton's VSM can be summarized in two ways: A) If $k = k_{\max}$ then LSI approaches the GVSM. Thus LSI improves Salton's method to IR by representing the data inter-dependence. Instead of assuming that the terms of a collection are independent, B) LSI attempts to improve the GVSM model of term correlations by dimensionality reduction. Where dimensionality reduction is performed by maintaining k dimensions that represent the highest term correlation in the term space. Thus LSI extends Wong GVSM by attempting to improve the model by creating a statistical model of the population correlation matrix via dimensionality reduction.

1.4 Dimensionality Reduction in Latent Semantic Analysis

Latent Semantic Indexing is related to other IR techniques such as multidimensional scaling (MDS), which use data visualization for exploring similarities or dissimilarities in data (Cox et al.,2001) and principal component analysis (PCA) , which reduce multidimensional data sets to lower dimensions for analysis (Jolliffe,2002). LSI is based on the singular value decomposition (SVD) of an input matrix, which will be discussed in

chapter 2. Given an $n \times p$ matrix A of rank r, the singular value decomposition of A is given in Equation 1.4.1:

(1.4.1) $$A = T \sum D'$$

In Equation 1.4.1, T is an $n \times r$ orthogonal matrix, $\sum$ is an $r \times r$ diagonal matrix, and D is an $m \times r$ orthogonal matrix. Where matrices T and D contain the left and right singular vectors of A respectively, while the main diagonal of $\sum$ contains the singular values, which are the positive square roots of A'A and AA'. The diagonal elements of $\sum$ reflects the amount of variance of the dimensionally reduced model from the original model (Hastie et al., 2001), (Rencher, 1995). Those diagonal elements of $\sum$ decrease in magnitude as i goes from 1 to rank k, this is demonstrated in Equation 1.4.2 where singular values follow a power law distribution hence the magnitude of singular values is related inversely and exponentially to the specified matrix rank k (Mihail et al.,2002),(Ding,2000).

(1.4.2) $$\rho_1 \geq \rho_2 \geq \rho_3 \geq ..... \geq \rho_r$$

Singular values decrease in magnitude as their rank increase, because they represent the amount of variance indicated by the corresponding dimensions from the full rank model. LSI suggests that we can improve information retrieval results by neglecting those singular values with small magnitudes (Deerwester et al., 1990).

Deerwester et al. found improvement over the VSM on several standard data sets. This can be achieved by removing dimensions with small corresponding singular values (Deerwester et al., 1990). Ding showed improvements in performance of 30% above traditional VSM-based systems on the ad hoc special retrieval task (Ding, 1999) (Ding, 2000). While Dumais applied LSI to several Text Retrieval Conferences (TREC) problems

(Dumais, 1992, 1993, 1994). Dumais research indicated a 31% improvement over keyword vector methods for the filtering task, and a 16% improvement for ad hoc retrieval (Berry et al., 1994).

Landauer and Dumais applied LSI to vocabulary learning problem. Their study results indicated that retaining approximately 300 dimensions yields the best accuracy for the vocabulary problem. They found that an LSI system is able to learn new vocabulary with accuracy over 50% (Landauer et al., 1997). Of particular interest about this study is the relationship between their system's dimensionality and its performance. Landauer and Dumais research indicates that when the number of dimensions (*k)* becomes much larger than 300, performance declines, this decline was interpreted as an evidence that the factors corresponding to small singular values contain essentially random noise distracters (Landauer et al., 1997), research results given by Ding (1999, 2000) and Story (1996) align with this hypothesis.

Research suggests that selecting the value of *k* (dimensions retained) is very important for good LSI performance. This indicates that a better LSI model should include factors whose corresponding singular values are large while discarding those that are small (Deerwester, et al., 1990).

Deerwester et al., called for the selection of an appropriate dimension as a very important factor for good information retrieval under LSI. However, moving from a low dimensionality of k = 1 to a moderately high dimensionality of k = 100 yields a 30% improvement in overall performance. Deerwester et al. says *"we are guided only by what*

appears to work best. What we mean by 'works best' is what will give the best retrieval effectiveness" (Deerwester, et al., 1990).

Additional interesting research question arise here which will be investigated in this research: What if we can't decide on the best dimensionality reduction parameter or technique in the unsupervised learning web environments, where noise and distracters effects cannot be neglected. In large information repositories a small change in the selected dimensionality might have a huge impact on overall system performance. Deerwester et al., method in selecting the reduced matrix dimension is common in many applications of LSI. However, the problem arises in practice where it is difficult to judge what does work best. In the case of Deerwester et al. or Landauer and Dumais, selection of k was performed by recourse to pre-classified data. All of these experiments make use of training data and test data that have been pre-classified, thus allowing the researchers to judge a given parameterization retrospectively by observing its accuracy on the test data. This approach is partly satisfying since most of the current IR systems do not have access to the relevance judgments that guide performance analysis used by Deerwester et al. In general, Deerwester et al. approach lacks a theoretical understanding for dimensionality reduction in IR systems implementation.

1.5 Effective Reduced Dimensionality Parameter

In studying dimensionality reduction parameters in LSI we encounter difficult questions on whether there is an existing optimal value for k. Jain et al., introduced the term of data set's intrinsic dimensionality which is also known as effective dimensionality (Jain et. al.1980). This term is also common in most literature that cover principal

component analysis where the intrinsic dimensionality is defined as a function of the multivariate probability density function responsible for the $n \times p$ matrix A, the intrinsic dimensionality of A is defined as the number of statistically uncorrelated variables in the probability density function of A, or the number of non-zero singular values (variances) in the population covariance matrix, the main observation in most studies is that those singular values for dimensions that exceed the matrix intrinsic dimensionality will tend to be small (Jobson,1991) (Rencher, 1995) (Jolliffe, 2002).

In general, intrinsic dimensionality is the minimum number of parameters that is necessary in order to account for all information in the data. Several techniques have been proposed in order to estimate the intrinsic dimensionality of a matrix. Major techniques will be discussed in chapter two.

1.6 Open Areas and Research Opportunities

Search engine results allow the user to view a document, navigate back to the search engine page and then based on the relevance judgment the user click on another relevant result, we conclude that this is not an ideal method, since hidden semantics of documents does not match user's level of knowledge to main concepts reflected by relevancy of results. Information retrieval techniques with latent semantic indexing try to limit the number of results returned to a user by reducing noise through dimensionality reduction. This can help accelerating relevancy process and direct users to relevant results. This activity supports user's cognitive model because domain knowledge is only contained at an abstract level.

In cognitive load theory, domain knowledge is critical in order to make an accurate relevancy judgment. The concept of cognitive load was presented by Miller (1956) where a

human's cognitive capacity for processing information was studied. Miller mentioned that "*The amount of information is exactly the same concept that we have talked about for years under the name of variance. The equations are different, but if we hold tight to the idea that anything that increases the variance also increases the amount of information we cannot go far astray*" Miller (1956).

It was concluded that working memory has a limited retention while other studies try to minimize cognitive load through interface design by recognizing human's working memory limitations. Studies in IR recognized that studying working memory limitations and capabilities may not be the only method of minimizing cognitive load. Beaulieu (1997) indicated that there is a need to study cognitive load to take account of the integration and interaction between the number and presentation of options, to add to this, I would like to refer to my research objective in finding a better structure of data collection to uncover concepts associations which are hidden as semantic properties, this will help answering questions such as: How much in the ranked list will users need to filter, to find all relevant documents?

Latent semantic analysis provides a measure for the similarity of meaning between words from text which are a close match to those of humans. Latent semantic analysis rate of absorption of knowledge from documents is similar to that of humans, and those results depend on the retained dimensionality of the representation. Latent semantic analysis performs similar to human-comparison. LSA performs well using representations that simulate multiple cognitive aspects that rely on word and passage meaning.

The similarity estimates obtained by latent semantic indexing are not simple correlations in usage, but depend on a powerful mathematical model capable of inferring

deep and strong relations (Latent Semantics), which are better approximates of human meaning-based reasoning and performance (Landauer,Foltz, and Laham, 1998). Latent semantic analysis reflects human knowledge since its scores overlap those of humans on standard vocabulary tests. Additionally latent semantic analysis simulate human word sorting and category judgments and accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay (Landauer,Foltz, and Laham, 1998). LSI can be used as a practical method for the specification of word meaning that provides measures of word-word, word-document and word-concept relations that are similar to several human cognitive aspects involving association or semantic similarity.

Intrinsic cognitive load is related to task difficulty, while extraneous cognitive load corresponds to task presentation. If intrinsic cognitive load is high, and extraneous cognitive load is also high, then problem solving may fail to provide correct solutions. When intrinsic load is low, then mental resources may remain to enable problem solving, even if a high level of extraneous cognitive load is required. Modifying the task presentation to a lower level of extraneous cognitive load will help maintain problem solving tasks if the resulting total cognitive load decreases to a level within the bounds of cognitive resources.

Literature review of research in dimensionality reduction indicted that no one to date has researched the effect of various information retrieval performance measures on overall retrieval performance when implementing a reduced matrix decomposition using LSA. As the dimensionality of data increases, query performance decrease and this is usually reflected and measured by the average system precision. This problem have been long

known as the “curse of dimensionality” was associated with much research in an effort to find better and more accurate techniques that process queries in large databases.

There is no consensus about the most effective method for estimating the best number of dimensions in LSI which results in better overall retrieval performance and that there is a need for research to be conducted on selecting the proper reduced matrix parameter in SVD which will yield improved overall performance. While this issue remains a challenging task, researchers have found that dimensionality reduction provides a better solution to information retrieval problems, which generally results in more relevant results and faster computational time, while giving reasonable accuracy and precision. An ideal dimensionality reduction technique has the ability of efficiently reducing data into a lower-dimensional representation, while maintaining the properties of the original data. Therefore it is desirable to find a technique that reduces dimensionality, while maintaining important information from the original model.

This research is going to contribute to reducing overall cognitive load through enhancing retrieval performance in terms of relevancy and better concept matching by finding the a better dimension that will yield improved overall search results in terms of relevancy, average search precision and recall while reducing the time it takes the user to find specific information, thus reducing the user level of uncertainty associated with the search process since the cognitive load will be reduced as users feels more confident that their information need can be answered.

Previous research performed on information retrieval systems using LSI has generally found improvements in search results, however, there are no studies which detail and evaluate the effect of selecting the reduced dimension on multiple performance measures. Studies in the

literature review indicated that LSI queries performance improve as the number of dimensions k increases, but this performance will decrease past a certain value of k. The value of k that enhances LSI performance is an open research issue, which will be studied in this research. One of the main objectives of my proposed research is to develop a new and improved model to investigate the effect of various dimensionality estimation techniques on overall search performance. This research will try to answer the following open questions in the implementation of LSI: What is the best method that enhances rank k approximation for the term-document matrix? Does a system using a weighted performance measures result in better overall performance? Does the weighted performance measures implementation provide an efficient LSI information retrieval technique than what we get by using full rank SVD?

CHAPTER TWO: LITERATURE REVIEW

2.1 Information Retrieval Systems

Information retrieval techniques search data repositories for documents that are relevant to users stated need via queries (Baeza-Yates et al., 1999) (Van Rijsbergen , 1979). Baeza-Yates, adopted a definition of basic IR vocabulary, by the use of the term *document* to denote a single unit of information and to describe text in digital form (Baeza-Yates et al., 1999). Queries are considered similar to documents, both mathematically and conceptually; or simply called "*pseudo-documents*". In older IR systems, documents contained only a few keywords, titles, or summaries of longer works (Cleverdon, 1967) (Luhn, 1961).

However due to the improved computing resources and the growth of electronic corpora such as the World Wide Web, documents in many newer IR systems contain a full reproduction of electronic texts. W. S. Cooper recommends intelligent information retrieval systems to borrow from machine learning, artificial intelligence, and linguistic research (Cooper, 1988). The volume and complexity of research into intelligent IR limits a general coverage of the subject. Instead, discussion will be limited to research in IR systems that build upon the vector space model.

2.2 The Vector Space Model (VSM)

Salton's vector space model (VSM) of IR characterizes retrieval in linear algebraic terms (Salton et al., 1975) (Salton et al., 1983). Under Salton's model, each document represents a vector in a *p*-dimensional vector space, where *p* is the number of indexing terms used. The location of the i^{th} document d_i along the j^{th} axis corresponds to the presence or absence of the j^{th} term in the i^{th} document. The simplest expression of the vector space model treats terms as binary data. Thus $d_{ij} \geq 1$ if the j^{th} term appears in the i^{th} document. Otherwise, $d_{ij} = 0$ (Salton, 1989). Table 2 contains a very small document collection about home cats and birds; Figure 2 depicts this data as points in a vector space. In this model, four documents are represented by two terms, cats and birds.

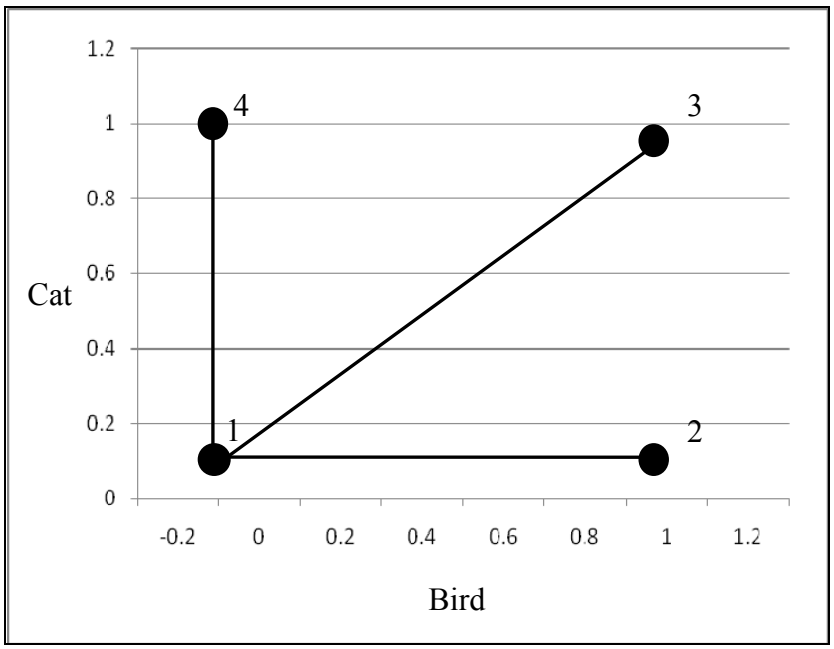


Figure 2: Home Cats and Birds data as vectors

Table 2: Home Cats and Birds Data

Document	Contents
1	Mans Best Friends
2	Feeding a Bird
3	Home Cats and Birds
4	Cat's lovers

The vector space shown in Figure 2 is defined as the space spanned by the rows of matrix A:

(2.2.1)
$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Matrix A shown in 2.2.1 is known as the *term-document matrix*; where the i^{th} column of A represents the i^{th} indexing term in document space. While the j^{th} row represents document j as a vector in term-space. Document number 1 contains neither indexing terms, and thus the model locates it at vector (0, 0) in A. Document 3 contains both birds and Cats, thus becomes (1, 1) in A. In vector space model, similarity between two documents i and j is defined as the inner product between the i^{th} and j^{th} document vectors, this is shown in Equation 2.2.2:

(2.2.2)
$$sim(i, j) = i.j = \sum_{m=1}^{t} i_m \bullet j_m$$

Normalizing the document vector to unit length gives the vector cosine shown in Equation 2.2.3.

(2.2.3) $$sim(i,j) = \cos_{ij} = \frac{i \bullet j}{\|i\|\|j\|}$$

A more common measure of similarity between document and query vectors is the cosine coefficient (Chowdhury, 1999), in which the similarity between a document in a collection d_j and query q is described by Equation 2.2.4

(2.2.4) $$sim(d_j, q) = \frac{{d_j}^T q}{\sqrt{\|i\|\|j\|}}$$

If we want to calculate the similarity between document 1 and 4 shown in matrix A (2.2.1) then $sim(1,4) = \frac{0+0}{\sqrt{0}+\sqrt{1}} = 0$, and $sim(3,4) = \frac{0+1}{\sqrt{2}+\sqrt{1}} = 0.71$, notice that under Salton's vector space model, documents 1 and 4 have no terms in common, while documents 3 and 4 share only one term, so we can say that documents 3 and 4 are closer together than documents 1 and 4. The query in Salton's vector space model is represented as a pseudo-document often denoted as q_i. Translating a query q_i into vector space model involves calculating $sim(q_i, d_i)$ then the model will try to presents results to the queerer ranked according to their similarity to q_i .

If we return to our birds and cats example, a query about birds or birds and cats will be transformed into a vector space representation in A as query vector q shown below:

$$q = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

VSM will then rank each document according to their similarity criteria shown in Equation 2.2.3

Vector space models have been of much importance due to its inter-document similarity representation. Salton's vector space model assumes that similarity is represented by geometric proximity (Salton, 1989). Salton assume that similarity is linear on the collection's indexing items. That is, vector space IR assumes that indexing terms are statistically independent. This assumption is proven to be false (Manning et al., 1999) (Oakes, 1998) (Cooper, 1988) (Cooper, 1991). Although it is unclear exactly how the assumption of term independence degrades the performance of IR systems (Losee, 1994).

Salton suggested the use of distinctions between individual terms based on their values to describe documents, where terms weight tend to be different based on several factors (Salton et al., 1988), Salton identifies two descriptors: *term frequency (tf)* for how many times the term appears in the document and *inverse document frequency (idf)* for how often the term appears in the information collection (Luhn, 1957). Luhn research suggested that most important terms in a document were those that are found with middling frequency (Luhn, 1955). Common terms such as "the", "in", "to" and "it" are over-represented in almost all English information repositories; their presence or absence provide little or no information about document relevance and aboutness discussed earlier. Many terms in a corpus will occur once or twice. These so-called terms provide too little information for useful text processing. Luhn (1955) suggests that terms that occur with mid-range frequency should be weighted when computing inter-document similarity. From this point Salton argues that any term weighting model should account for term frequency.

The notion of inverse document frequency (*idf*) was introduced by Karen Sparck Jones. According to Sparck Jones it is not sufficient to consider a term's global frequency (*tf*) when estimating its usefulness for discrimination. Analysis of a term's distribution across documents should supplement *idf* analysis (Sparck Jones, 1972). This consideration stems from the possibility that a term could be quite common, but present in only a small subset of a corpus' documents. A purely *tf*-based model would degrade such terms due to its common appearance, although its concentrated distribution suggests that it could serve as a useful marker for a subclass of document (Sparck Jones, 1979). Thus the *idf* factor as Salton mentioned, "*Varies inversely with the number of documents n to which a term is assigned in a collection of N documents. A typical idf factor may be computed as* $Log\frac{N}{n}$ *"* (Salton et al., 1975), thus Salton was able to develop an IR weighing scheme for term discrimination which assumes that best terms should have high term frequencies but low overall collection frequencies. To estimate terms discrimination value, Salton used the product of *(tf)* and *(idf)* (Salton et al., 1988). Although this term weighing criteria have been criticized because of insufficient theoretical foundations (Bookstein et al., 1975) (Cooper et al., 1978), this term weighing criteria was popular in many IR research (Bollacker at al., 1998) (Joachirns, 1998) (Prey, et al., 2001). In general Salton's vector space model imagines that all terms are equally important, and that their presence or absence with the frequency of their repetition determines the conceptual content of a document. So that, for the term discrimination model, not only does it matter how many times a term appears in a document, but it is also important to know how many documents contain the term. In this case we are reducing the model from a space vector in *p-space* to a vector in *k-space*, where $k < p$.

While analyzing each term's distribution (*idf*) across documents, the suggested model accounts for documents inherent features, suggesting that those terms that are largely used in a small group of documents will be strong indicators for retrieval purposes.

Many words in data repositories are only slightly useful for information retrieval systems. Stop-lists (Baeza-Yates et al., 1999) (Salton et al., 1983) (Salton et al., 1989) were created for removing high-frequency terms (Noise) which adds no useful information. Likewise, the use of stemming (Porter, 1980) can reduce the number of indexed terms by mapping variants of a stem down to a single root. Researchers would benefit from stemming by eliminating these document features that adds noise into the document ranking process. Salton suggested that if our weighting model is up to the task, we may derive the *k* most important features in a collection by ranking the terms by *idf* weight and keeping top *k* ranks.

Influenced by Cooper's results (Cooper,1991), Salton included the effect of term dependencies and correlations to the Vector Space Model in a number of ways, of them: (A) Generating sets of related terms by observing co-occurrence in data from online corpora (Lesk, 1969),(Van Rijsbergen, 1977),(Church et al., 1990). (B) Identifying common word phrases and considering them indexing features similar to individual words (Sparck Jones et al., 1984). (C) Use of online thesauri (Amsler, 1984) (Sparck Jones et al., 1984) (Fox, 1980), (Fellbaum, 1998). (D) Development of knowledge bases and logical relations among indexing terms (Croft, 1986) (Croft, 1987).

The main objective of relevance feedback adopted by Rocchio technique (Rocchio, 1971) is to construct an optimal query q_{opt} by studying retrieved documents in the collection C_r for a given query q .

Rocchio method start by assuming that we have a complete knowledge of the relevance values for a query q for every document in our collection. This is given in Rocchio Equation 2.2.5, where the optimal query is a weighted sum of relevant and non-relevant document vectors, with the weights depend on the size of C_r in relation to the size of the collection. Symbols used in Rocchio relevance feedback is shown in Table 3.

(2.2.5) $$q_{opt} = \frac{1}{|C_r|} \sum_{\vee d_j \in C_r} d_j - \frac{1}{N - |C_r|} \sum_{\vee d_j \in C_r} d_j$$

Table 3: Symbols used in Rocchio Relevance Feedback

Symbol	Meaning
D_r	Set of relevant documents among retrieved documents
D_n	Set of non-relevant documents among retrieved documents
C_r	Set of all retrieved documents
$\|D_r\|, \|D_n\|, \|C_r\|$	Number of elements in each set of documents
α, β, γ	Constant Parameters

Because we do not have access to the requisite sets of relevant and non-relevant documents, the final query vector under Rocchio technique is formed by Equation 2.2.6. The objective of relevance is to manifest the relationships between terms (Ide, 1971).

(2.2.6) $$q_m = \alpha q + \frac{\beta}{|D_r|} \sum_{\vee d_j \in D_r} d_j - \frac{\gamma}{|D_n|} \sum_{\vee d_j \in D_n} d_j$$

Rocchio relevance feedback constructs an ideal solution vector q_m as shown in Equation 2.2.6, which is the best linear approximation of q_{opt} as q_m which maximizes the similarity

(minimize distance) between the query and the center of the set of relevant documents while maximizing its distance from the center of the set of non-relevant documents, this will guarantee an optimal query generation. Salton (1989) describes three main advantages to using similarity coefficients between query and document vectors:

1) Documents can be arranged in descending order of similarity.
2) The number of documents retrieved can be limited to the most similar documents.
3) Documents located early in the list of retrieved documents might be the most useful documents according to their relevance to the query.

2.3 Latent Semantic Indexing (LSI)

Research on term-based information retrieval shows the side effects of undue cognitive burden placed upon end-users interested only in abstract concepts rather than in specific and accurate technical words (Furnas, Get al., 1987) (Newby, 2001). Information retrieval cognitive research suggests that development in IR should account for psychological developments in the cognitive sciences, Newby suggests a *"computerized representations of data sets as found in document collections which are compatible with human perception of the data sets"* (Newby, 2001). Newby mentioned two useful statements, which will be followed in this research. The first statement is that information space domain is the set of concepts and relations between them held by a computer system. And the second statement is that information spaces are comprised of words, documents, and the relations among them. Based on this, cognitive space is defined as the set of concepts and relations between them held by human knowledge. Although it is difficult to identify the fundamental components of cognitive spaces in human's knowledge and

experience, psychological research finds a high degree of similarity among psychometric analyses of individual linguistic association (Wittgenstein,1953), (Rosch,1975),(Rosch, et al.,1976).

The basic assumption behind latent semantic indexing (LSI) is that term correlation in information retrieval reduces searchers cognitive burden. LSI was created to address the gap between information spaces and cognitive spaces so as to improve VSM representation to accommodate for the error of term independence (Landauer et al., 1997), (Landauer et al., 1998), (Foltz et al., 1998),(Gardenfors, 2000),(Landauer, 2002),

LSI addresses two main problems in IR: *Polysemy*, or the problem that many words have more than one meaning, and that those meanings are obtained from the context in the documents collection. And *Synonymy*, or the problem that there are sometimes more than one way of describing the same object. Synonyms tends to decrease the recall retrieval performance of IR systems (Deerwester et al., 1990).

LSI implements dimensionality reduction, hence the latent semantic space which is created in LSI has fewer dimensions than the original space (Manning et al., 1999). LSI based systems are able to match and find terms which do not appear in a document. Thus documents located in a similar space of meaning will be retrieved. Latent semantic indexing use statistical modeling to improve the representation of terms and documents by deriving a low-rank approximation, A_k of the term-document matrix, A where A_k provides the best least-squares rank-*k* of A. In projecting the information space onto a low rank A_k , LSI achieves two main benefits over the standard vector space model: The inclusion of terms dependence, and dimensionality reduction. (Deerwester, et al., 1990), (Berry et al., 1994), (Husbands et al., 2000).

Deerwester et al. referred to the points of weaknesses in information retrieval methods in that the words searchers use are not always similar to those by which the information they seek has been indexed, Deerwester et al. referred to the problem of *synonymy* and *polysemy* (Deerwester, et al., 1990).

Synonymy affects searchers when searching with different terms in a query than what an author or indexer used in a relevant document. Thus retrieval systems might fail to deliver documents about homes when presented with a query about houses or queries about cars when presented with queries about automobiles. Information retrieval performance is also downgraded due to polysemy because natural language terms tend to have multiple different meanings, the term can imply quite different topics in different contexts. LSI relies on statistical modeling which approximates the dynamics of a variable and stochastic system. Neter et al. mentioned that statistical models contain two components (Neter, et al., 1996): (A) Functional Element, with which the model expresses the relations among system variables as a mathematical function. And (B) Stochastic element, which assume that the behavior of the system is non-deterministic, but rather that its dynamics is in part governed by a set of probability distributions.

According to Neter, et al. (1996), a mathematical model describes a system deterministically, thus we may construct as an example a model to calculate a family monthly payments based on the number of services or purchases that they have in a specific month. Such a model defines two kinds of variables, a dependent (response) variables based on given information of other variables in the system and an independent variables (predictors) that provide information by which we predict the value of a dependent variable.

However statistical models are different than mathematical models due to their semi-deterministic nature (Bhattacharyya et al., 1977).

Statistical models are constructed empirically by following three general steps (Jobson, 1991) (Neter, et al., 1996): (1) choosing the family of functional relations which describe the system behavior. Mainly the family of linear functions is widely used in information retrieval due to their mathematical descriptive power (McCullagh, et al. 1989), (Cherkassky, et al., 1998), (Hastie. et al., 2001). (2) Identifying the probability distribution that governs the variability of the system. (3) Parameterizing the model function.

LSI apply linear regression as one of its main modeling techniques, Story (1996) provided a detailed discussion of the relation between information retrieval and linear regression (Story, 1996). A simple linear regression model is shown in Equation 2.3.1, where y_i is the i^{th} response, β_o and β_1 are fitted parameters, x_i is the i^{th} observation, and ε_i is the i^{th} error term.

(2.3.1) $$y_i = \beta_\circ + \beta_1 x_i + \varepsilon_i$$

Least-squares method is used to solve the regression model. In solving for this we choose those regression coefficients that minimize the squared error between the observed data, and the predictions at each observation, thus we find a fitted value for the response, $y_i = \beta_\circ + \beta_1 x_i$, where the sum of squared errors (SSE) is shown in Equation 2.3.2 (Forsythe, et al., 1977) (Neter, et al., 1996)

(2.3.2) $$SSE = \sum (y_i - \hat{y}_i)^2$$

In fitting the least squares estimate of the model we find the parameters that minimize SSE, β_o and β_1 or the parameters that minimize the residual deviance of the model. (Fisher, 1974) (Jobson, 1991) (Neter, et al., 1996). Additionally, least-squares estimate of the regression coefficients is shown Equation 2.3.3 where the covariance is given by $X'X$ (Jobson, 1991), (Rencher,1995).:

$$\hat{\beta} = (X'X)^{-1} X'y \tag{2.3.3}$$

To measure the variance that is captured by the regression model we calculate the coefficient of determination R^2 which is a measure of the descriptive power of the model as shown in Equation 2.3.4 (Burnham et al., 1998).

$$R^2 = \frac{\hat{\beta}x'y - n\bar{y}^2}{y'y - n\bar{y}^2} \tag{2.3.4}$$

The main objective of LSI is to generalize from observations, this is the result we get from implementing linear models approximations, according to this we can simulate LSI process as a series or linear regression processes. In this sense LSI tries to build relations that were neglected in VSM that accounts for term independence.

Rencher (1995) described principal component analysis as a method that tries to maximize the variance of a linear combination of a variables by searching for the optimal dimension that maximize data spread, this is meant to organize information according to the main topic, a problem described as an eigenvalues-eigenvector problem (Rencher,1995), (Strang, 1998), (Jobson, 1991). PCA assumes that dimensionality reduction helps overcome

sampling error. According to Rencher, PCA align the principal components or variances from the largest (sample) variance to the smallest sample variance (Rencher, 1995). Thus by retaining only the first $k < p$ principal components we achieve the best rank-*k* approximation of the covariance matrix, in the least squares sense.

2.4 Singular Value Decomposition (SVD)

Latent semantic indexing use a low rank approximation of the original data matrix A by adopting the use of singular value decomposition (SVD), a least-squares matrix factorization method from linear algebra (Golub, et al., 1989), (Forsythe, et al., 1977), (Berry et al., 1994), (Strang, 1998) .The singular value decomposition of matrix A $(n \times p)$ of rank r is shown in Equation 2.4.1

(2.4.1) $$A = T_{[m \times r]} \sum_{[r \times r]} D_{[r \times n]}$$

In Equation 2.4.1 T and D are orthogonal matrices, where T is $m \times r$, with columns t_i containing the left singular vectors of A. D is an $r \times n$ matrix with columns d_i; referred to as the right singular vectors of A. Matrix $\sum$ is an $r \times r$ diagonal matrix, with the diagonal elements $\rho_1 \geq \rho_2 \geq \rho_3 \geq ..... \geq \rho_r \geq 0$ called the singular values (Deerwester, 1990) (Berry et al., 1994) (Hastie et al., 2001).

The matrix of singular values $\sum$ acts as a reference of the amount of variance described by each factor *k* in the derived factor space (Jobson, 1991). This property is useful when selecting singular values (variances) to retain during dimensionality reduction. SVD is used to derive a least-squares approximation of A, as shown in Equation 2.4.2, where all

term-document similarities are approximated by the results of this model with reduced dimensions (Deerwester, 1990):

(2.4.2) $$\hat{A}_k = T_k \sum_k D_k$$

In Equation 2.4.2 T_k contains the fist k columns of T, $\sum_k$ contains the first k rows and columns of $\sum$, and D_k contains the first k columns of D. Thus the similarity between two documents represented as vectors d_i and d_j is the inner product between the i^{th} and j^{th} rows of D_k. A query q is added as an ad hoc document and projected as shown in Equation 2.4.3 (Berry et al., 1994).

(2.4.3) $$q_k = qT_k \sum_k$$

We calculate similarity between queries and each document in the corpus by applying Equation 2.2.3 to find $sim(q_k, d_k)$ where d_k is the i^{th} row of D_k .

One of the major strengths of LSI is its ability to identify topical clusters of terms and documents. LSI is considered an extension to Wong generalized vector space model, since it augments standard vector space model to include an analysis of the correlational structure of data (Wong et al., 1987), (Jiang et al., 2000), an example of SVD term document structure of rank five is shown in Figure 3.

Deerwester et al. stated that choosing a dimensionality that indicates the correlational structure of the population from which a data sample is drawn is an open problem in the literature (Deerwester et al., 1990). Deerwester et al., indicated that the representation of large document collections will require more than a collection of underlying independent

concepts which manifest the importance of the amount of dimensionality reduction or selected dimensions k to effective implementation of LSI (Deerwester et al., 1990).

$$T = \begin{pmatrix} n_{11} & . & . & . & n_{15} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ n_{51} & . & . & . & n_{55} \end{pmatrix}$$

$$\sum = \begin{pmatrix} n_1 & 0 & 0 & 0 & 0 \\ 0 & n_2 & 0 & 0 & 0 \\ 0 & 0 & n_3 & 0 & 0 \\ 0 & 0 & 0 & n_4 & 0 \\ 0 & 0 & 0 & 0 & n_5 \end{pmatrix}$$

$$D = \begin{pmatrix} n_{11} & . & . & . & n_{15} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ n_{81} & . & . & . & n_{85} \end{pmatrix}$$

Figure 3: Example of SVD term-document structure or rank five

In order to process a query with multiple keywords in latent semantic indexing we need to represent each term and document as a vector in k dimensional space (we would like to use the reduced matrix dimension or intrinsic dimensionality) that will improve overall

performance. A query will be treated just as a document which appears as a set of keywords. Thus a query or "pseudo-document" will be represented as the weighted sum of component term vectors.

To get a set of potential relevant documents, the pseudo-document (query) formed from multiple keywords is compared against all documents using the euclidian distances or vector cosines by multiplying the corresponding values of each query terms by the documents weighted term frequency values, we select those values with the highest cosines, that is the nearest vectors with high corresponding documents similarities, to be returned as relevant documents. Generally a limit or threshold is set for the closeness of documents and all those documents above the threshold or within the *n* closest are returned. This cosine measure is an indicator of similarity to predict human relevance judgment regarding similar concepts in a text collection; in addition to the effects of dimensionality reduction to improve information retrieval relevancy measure and reducing overall user cognitive load. To illustrate this, the following example is provided for two imaginary groups of documents from a collection in Computer Science (CS) and Human Factors (HF). This collection is summarized in table 4 and 5

Table 4: Example for two groups of documents from an imaginary collection in computer science (CS) and human factors

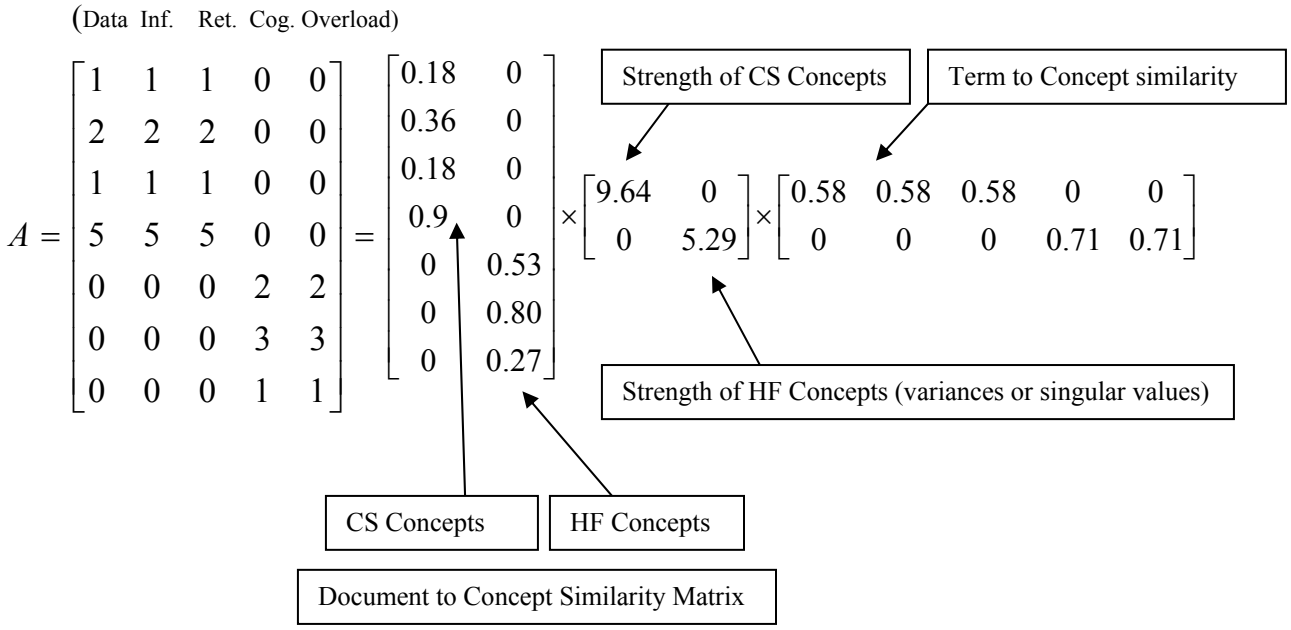
Type	Terms / Document	Data	Information	Retrieval	Cognitive	Overload
CS	Doc-CS-1	1	1	1	0	0
CS	Doc-CS-2	2	2	2	0	0
CS	Doc-CS-3	1	1	1	0	0
CS	Doc-CS-4	5	5	5	0	0
HF	Doc-HF-1	0	0	0	2	2
HF	Doc-HF-2	0	0	0	3	3
HF	Doc-HF-3	0	0	0	1	1

Term document matrix A will be decomposed into:

$$\mathbf{A}_{[n \times m]} = \mathbf{T}_{[n \times r]} \sum_{[r \times r]} (\mathbf{D}_{[m \times r]})^T$$

Table 5: Term document matrix decomposition details

A:	n x m matrix (*n documents, m terms*)
T:	n x r matrix, document-to-concept similarity matrix (*n documents, r concepts*)
$\sum$:	r x r diagonal matrix with diagonal elements representing 'strength' of each concept (positive singular values representing the variances), and sorted in descending order ($\sum$ i: *strength of each 'concept'*) (*r: rank of the matrix*)
D:	m x r matrix, term-to-concept similarity matrix (*m terms, r concepts*)



To search for queries we need to represent query vectors into 'concept space' as an example consider Query 1 with two keywords which searches for "Information Retrieval"

(Data , Inf. , Ret., Cog. ,Overload)

$$Query\ 1 = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Query (similarity to concepts) = (Query1) x D (Term to Concept matrix)

$$Query\ 1 = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{bmatrix} 1.16 & 0 \end{bmatrix}$$

Query Vector — Term to Concept similarity

HF Concepts similarity to Query 1

CS Concepts similarity to Query 1

$$Query\ 2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{bmatrix} 0.58 & 0 \end{bmatrix}$$

Query Vector Term to Concept similarity

HF Concepts similarity to Query 2

CS Concepts similarity to Query 2

We notice that documents with keywords of ('information', 'retrieval') will be retrieved by query ('data'), although it does not contain 'data'!! This highlight the importance and benefit of dimensionality reduction for retrieving relevant documents, when we have large document collections with thousands of terms from different concepts, using similarity generalizations to queries will add great value to relevancy measure and reduce noise which will distract users.

2.5 Term Weighting

Terms weighting assign terms which are more important a higher value than less important terms. Summing the number of times each term appears in a document is the most used and simplest term weighing technique. The use of term weighting usually results in better ranking (Frakes et al., 1992). Equation 2.5.1 is a weighting scheme that consists of three components, where a_{ij} is the $ij-th$ element of term document matrix A:

(2.5.1) $$a_{ij} = g_i\, t_{ij} d_j$$

In Equation 2.5.1, g_i is a global weight which is applied to all non-zero occurrences of term i (all values of row i). t_{ij} is the local weight for term i in document j (Kolda, 1997). d_j is a normalization factor which may be required as larger documents will tend to receive a higher similarity coefficient, due to higher term frequencies. Kolda (1997), Salton and Buckley (1997b) provided a more comprehensive list of weighting formulas.

2.6 Stop Lists

Candidate terms are usually compared against a stop list during the automatic indexing of documents. Stop list is a list of very common words (e.g. "the", "an".., etc.). Those terms appear in most documents and will be removed when they show up frequently. The advantages of using a stop list is that less storage space is required and that high frequency terms are removed from both the query and the term matrix which means faster retrieval . The disadvantage of using a stop list is that search phrases might require words from the stop list. The standard Stop list used in many IR studies is the list used by *SMART* program, which contains 429 terms. This list is shown in Appendix A.

2.7 Stemming

Stemming is a morphological collapsing of word variants into a single root. For example, `*Simulate*', `*Simulation*' and `*Simulated*' will all have the same root `*Simulate*'. Jurafsky et al. indicated that stemming needs to be applied to the keyword matrix and to the query in order to be effective. The advantage of stemming is that a query on the keyword `*Simulation*' will be stemmed to `*Simulate*' before start searching for it in a document collection and will retrieve documents which also use the keyword '*Simulated*' and

`*Simulating*' (Jurafsky et al., 2000). Frakes (1992) found that there was little difference in IR system performance between stemming methods. The disadvantage of stemming is that it can return terms which have stemmed to the same root, but are not related to the query. Research on the benefits of stemming is inconclusive, although stemming generally doesn't degrade retrieval effectiveness (Frakes, 1992). The Porter stemmer (Porter, 1997) is the most commonly used stemming algorithm, due to its simplicity.

2.8 Reduced Dimension of the Singular Value Decomposition

Research in LSI suggests that dimensionality reduction removes the noise from the term document matrix representation. Dimensionality reduction projects the term document matrix r into an orthogonal subspace or a lower rank k where $k << r$. However, the reduced dimensionality k is not fully understood in applications of LSI, and the source and character of the noise is difficult to understand and verify. Additionally the actual error distribution of these models is not clear (Manning et al., 1999) (Husbands et al., 2000). Results indicate that without a complete understanding of these models, ignoring remaining dimensions ($r - k$) introduce risk on inaccurate models (Ding, 2000).

After Deerwester et al. proposed their approach in information retrieval using LSI; researchers noticed that properly parameterizing the representational dimensionality of the model is a vital for information retrieval accuracy and precision. Deerwester et al. mentioned that the reduced dimensionality parameter is crucial to successful application of LSI (Deerwester et al., 1990), noting a 30% improvement in average precision as they changed k from 1 to 100 on the Medline data collection. Setting the model dimension to very low values will deprives the model from important descriptive power to perform

consistent information retrieval, Deerwester et al. indicated that setting the model dimensions to low number of factors, k = 100, yields good overall performance.

Landauer and Dumais write, "*Using too many factors (for LSI representations) also resulted in very poor performance*" (Landauer et al., 1997). It was indicated that setting $k = 1$ leads to accuracy slightly below 16% on a synonym learning test. In the region of $k = 300$, Landauer and Dumais report accuracy above 50%. As they increase k, letting it approach the full matrix dimensionality of their data collection, accuracy dips back to the 15% level Landauer and Dumais test the validity of this strong non-monotonic relation between the number of dimensions and the accuracy of simulation, by recourse to a statistical hypothesis test, noting a *p*-value below *0.0002* (Landauer et al., 1997).

In practical implementation, researchers tried to approximate k_{opt}. Deerwester et al. indicated a region with a corresponding optimal dimensionality, where k_{opt} was selected by approximation. Deerwester et al. wrote "*We have reason to avoid both very low and extremely high numbers of dimensions, In between we are guided only by what appears to work best. What we mean by 'works best' is ... what will give the best retrieval effectiveness*" (Deerwester et al., 1990). Landauer and Dumais indicated that identifying k_{opt} for a given corpus is a complex issue that must be addressed in future research (Landauer et al., 1997).

Landauer and Dumais work formalizes a pattern that is encountered often in applied LSI research: for data collections there is a region of optimal dimensionality less than the full rank of the dataset. Reducing the matrix dimension and setting k a value below this region deprives the system of important descriptive power, while setting a value of k that is too high appears to over-fit the model (Landauer et al., 1997), this means that the model will

learn additional term-document relations, which reduce LSI ability to predict correct term-document associations.

Manning et al. mentioned a region of optimality with regard to parameterizing k in LSI models. These observations suggest that observing the performance of an LSI system at various levels of k gives an indication about the intrinsic dimensionality of a data collection (Manning et al., 1999). Ding mentioned that adding factors to an LSI model quickly improves performance until a certain threshold is reached. After this region of optimality, performance decreases as one adds more singular vectors (Ding, 1999) (Ding, 2000).

Dumais indicated the need for more dimensionality representational details than what a 100 dimension can afford to be able to represent a 742,331 document by 104,533 term matrix, Dumais derived a smaller matrices by document sampling. Analyzing these sampled matrices by SVD, Dumais choose values of k ranging from 200 to 300 (Dumais, 1993). These results suggest that while larger corpora demand more factors, this increase is not linear. On the other hand, small collections might perform well under k = 5% of the number of documents, while representing a large corpus may only require k =0.005% of the number of documents (Dumais, 1993).

Previous approaches which estimate matrix intrinsic dimensionality relies on pre-classified test data to define a well-constructed model, this is common in IR evaluation, however, an open research question is to find a model goodness of-fit that is applicable to the unsupervised learning environment such as large data collections and the World Wide Web information collections. Hofmann (2001) criticized the normality assumption which is introduced by least-square method. The method of least-squares minimizes the model's

squared error $(x-\mu)^2$ shown in Equation 2.8.1.1. This will provide a solution based on the assumption of normality.

(2.8.1.1) $$n(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$$

Information retrieval research indicated that term-document matrices are non-normal in their distribution. A well-known research holds that term count data tend to follow a Zipf-like distribution (Manning et al., 1999) (Jurafsky et al., 1999) (Zipf, 1929) (Mihail et al.,2002) (Efron, 2003). The Zipf distribution is called the power law model, which suggests that the rank and frequency of terms in a data collection will be inversely and exponentially related. Thus many terms occur once or twice; while only a few terms occur often.

In order to help solving the problem of dimensionality reduction in IR, Hofmann proposed a probabilistic latent semantic analysis (PLSA) model (Papadimitriou et al., 1998), (Hofmann, 1999) (Hofmann, 2000). In PLSA model, the k factors derived by LSI are noticed to correspond to the mixture of various components. As such, "*the mixing proportions in PLSA substitute for the singular values of the SVD in LSA*" (P.184, Hofmann, 2001), this model finds the best retrieval performance by using a linear combination of models, each fitted with a different k-value. Thus while LSI may violate certain assumptions considered in the least-squares model, its mathematical simplicity (as a least-squares method) and its good performance, contribute to its advantages. Ding proposed a "dual probabilistic model" similar to the maximal likelihood model, findings were that LSI is the optimal solution of the model. Equation 2.8.1.2 demonstrates the maximum likelihood for a k-dimensional model in LSI (Ding, 2000) (Efron, 2003):

(2.8.1.2) $$l_k = \sigma_k{}^2 + \sigma_k{}^2 + \ldots + \sigma_k{}^2$$

In Equation 2.8.1.2 , σ_k is the k^{th} singular value of A. Adding weak singular vectors increases the model likelihood by a small amount. Through using Ding's model we can acquire a precise conclusion of the contribution of each singular vector to the overall representation. Ding indicates that the contribution, or the statistical significance, of each LSI dimension is nearly the square of its singular value (Ding, 2000). Thus, each factor's statistical significance is represented by a quadratic relation to the magnitude of the corresponding singular value, where small singular values correspond to very small contribution, and this means negligible improvements in model likelihood. Overall, Ding's model does not provide a solution to the problem of selecting k for an LSI model.

Rencher (1995) indicated the importance of inter-variable correlation among data collections. In this study, Rencher concluded that the degree of dimensionality reduction required for best performance is proportional to the degree of correlation among the matrix variables. This indicates that the highest few singular values will capture the system variance. According to Rencher, if the variables are highly correlated, then the reduced dimension is much smaller than the original matrix rank; only the first few singular values will have large values that affect the predictability of the LSI model, while on the other side there is no need for dimensionality reduction if the correlations among the variables are small, since matrix intrinsic dimensionality is close to the original matrix size (Rencher, 1995) (Efron, 2003).

One of the suggested methods to calculate the value of dimensionality reduction was based on a hypothesis testing to find if the $p-k$ smallest singular values are equal, this methods is called Bartlett's test of isotropy for dimensionality reduction. In this method we

test the null hypothesis that $H_O : \lambda_{k+1} = \lambda_{k+2} = \lambda_{k+3} = ... = \lambda_p$, thus if the null hypothesis is true, we conclude that there exist no dimensional subspace in the $p-k$ singular values, while if any of the $p-k$ singular values is significantly less that $\bar{\lambda}$ then there exist a reduced dimensionality at this point. Based on this assumption we either reduce the matrix dimension or don't reduce dimensionality at all (Krzanowski, 1988). Bartlett's test of isotropy starts by calculating the average of the last $p-k$ singular values as shown in Equation 2.8.1.3

(2.8.1.3) $$\bar{\lambda}_k = \sum_{i=k+1}^{p} \frac{\lambda_i}{(p-k)}$$

To find Bartlett's test statistic we use Equation 2.8.1.4 where n is the number of data observations, and the test statistic u is approximately χ^2-distributed.

(2.8.1.4) $$u = (n - \frac{2p+11}{6})(k \ln \bar{\lambda}_k - \sum_{i=k+1}^{p} \ln \lambda_i)$$

Thus according to equation 2.9.1.4 we reject H_O if $u \geq \chi_{\alpha,v}^2$ where $v = \frac{1}{2}(p-k-1)(p-k+2)$. Bartlett test of isotropy continue to find k_{opt} by testing $H_{02} = \lambda_{p-1} = \lambda_p$, then if H_{02} gives high confidence level we test $H_{03} = \lambda_{p-2} = \lambda_{p-1} = \lambda_p$ and we continue until we stop at k_{opt} at which no sufficient confidence level that the last $p-k$ singular values are equal (Anderson, 1984) (Jobson,1991).

Kaiser-Guttman technique or Eigenvalue-one criterion is the most popular method for dimensionality reduction and for identifying significant principal components (Guttman, 1954). Kaiser-Guttman technique retains those factors with corresponding singular values

greater than the average of all the singular values, where the k^{th} singular value is the amount of variance described by the k^{th} principal component. Thus we need to include all singular values greater than the average or include all correlation matrix singular values which are greater than $\overline{\lambda}$.

Retaining all singular vectors whose corresponding singular values are greater than $\overline{\lambda}$, means keeping those factors that describe more variance than the average observed variable in the original data set. However, if documents are orthogonal, indicating independence, then all singular values will be similar and Kaiser-Guttman technique returns a full dimensionality model. One of the drawbacks of Kaiser-Guttman technique is the assumption that population parameters are used, and not sample statistics (Guttman, 1954). However, in common practice we work with samples, not population parameters. Problems in applications of Kaiser-Guttman rule arise, because Guttman's procedure does not recognize the distinction between the observed correlation matrix and the population correlation matrix.

A re-sampling procedure called Parallel Analysis (PA) was introduced to help include the effect of the sample correlation matrices (Horn, 1965) (Subhash, 1996). PA generates many $n \times p$ data sets A^* from the normal distribution with a mean vector of the original matrix A and the identity matrix I_p for the covariance matrix, in this way Horn (1965) introduced sampling error to the model. We proceed by averaging the singular values of $\lambda_1^*, \lambda_2^*, \lambda_3^*, ...\lambda_p^*$ across all samples to get $\overline{\lambda}_p^*$, a vector of the singular values generated from the independent variables, those values are compared to observed data to find k_{opt}. Parallel analysis is considered as an improvement to Kaiser-Guttman because it considers

that we are analyzing data with a sample size of $n < \infty$. Efron (2003) introduced Amended Parallel Analysis (APA) technique based on parallel analysis and conducted several simulated tests on APA with good results. The number of resampling iterations required to get k_{opt} is an open research question, setting the number of samples to 100 have been a common approach in many studies (Efron, 1993).

The percentage of variance technique was introduced by Dillon (Dillon, et al., 1984), in order to chooses a limiting point, m , that represent the proportion of observed variance that the final model have to introduce; in this technique the fewest singular values sufficient to account for *m* percentage of the variation among the original data is considered, commonly 90~95% (Rencher, 1995) (Jackson, 1993). Thus we calculate the percent of variance captured by the first k singular values through Equation 2.8.1.5.

(2.8.1.5)
$$m_k = \frac{\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_p}$$

Some studies indicated the suitability of *m* around 85% for large and complex datasets which requires more dimensionality reduction. It have been indicated that both Kaiser-Guttman and the percent of variance techniques have been used extensively in applied statistics and had much popularity in dimensionality estimation for various software packages (Jolliffe, 2002).

Maaten et. al. (2007) described a collection of various dimensionality estimation techniques and included implementations of 27 techniques for dimensionality reduction. Additionally, there is a description of 6 intrinsic dimensionality estimators and functions for out-of-sample extension and data generation (Maaten, 2007).

As discussed earlier in chapter one, there are two groups of intrinsic dimensionality estimation techniques, (A) estimators based on the analysis of local characteristics of the data and (B) estimators based on the analysis of global properties of the data. (Maaten et. al. 2007). Local intrinsic dimensionality estimators are based on the observation that the number of data points covered by a hyper-sphere around a data point with radius r grows proportional to the matrix dimensionality r^d, where d is the intrinsic dimensionality of the data around that data point. The intrinsic dimensionality d can be estimated by measuring the number of data points covered by a hyper-sphere with a growing radius r. There are three local intrinsic dimensionality measures, the correlation dimension estimator, the nearest neighbor dimension estimator, and the maximum likelihood estimator (Levina, et. al., 2004). The correlation dimension estimator uses the intuition that the number of data points in a hyper-sphere with radius r is proportional to r^d by computing the relative amount of data points that lie within a hyper-sphere with radius r. The nearest neighbor estimator is similar to the correlation dimension estimator; however, it computes the minimum radius r of the hyper-sphere that is necessary to cover k nearest neighbors. The maximum likelihood estimator estimates the number of data points covered by a hyper-sphere with a growing radius by modeling the number of data points inside the hyper-sphere as a Poisson process (Levina, et. al., 2004) (Burges, 2004) (Maaten et. al. , 2007).

Global intrinsic dimensionality estimators consider the data as a whole when estimating the intrinsic dimensionality. There are three global intrinsic dimensionality measures; the Eigenvalue-based estimator, the packing number estimator, and the geodesic minimum spanning tree estimator. The Eigenvalue-based intrinsic dimensionality estimator performs PCA on the high-dimensional dataset and evaluates the Eigenvalue corresponding to the

principal components (Fukunaga, et. al., 1971). (Maaten et. al. 2007). The packing numbers intrinsic dimensionality estimator is based on the intuition that the number of hyper-spheres with radius r that are necessary to cover all data points with radius r is proportional to r^{-d}, in other words, the packing numbers intrinsic dimensionality estimator is the maximum number of data points that can be covered by a single hyper-sphere with radius r. The geodesic minimum spanning tree (GMST) is the minimum spanning tree of the neighborhood graph defined on the dataset. The length function of GMST is considered to be the sum of the Euclidean distances corresponding to all edges in the geodesic minimum spanning tree (Maaten et. al. 2007). Burges (2004) analyzed several geometric methods for feature selection and dimensional reduction by dividing the methods into projective methods (e.g. PCA) and methods that model the manifold on which the data lies (e.g. MDS). Figure 4 demonstrate the taxonomy of intrinsic dimensionality estimation techniques described in previous dimensionality reduction research. Table 6 provides a summary of published works that consider dimensionality reduction techniques in information retrieval.

Literature review indicates that there is no consensus on the most effective method for estimating k_{opt} in LSI and that there is no research conducted on finding the parameter of the reduced matrix dimensionality that will satisfy multiple performance measures.

The matter of dimensionality selection remains an open research area and important problem. Hofmann mentioned in the context of fitting LSI models, "*deriving conditions under which generalization on unseen data can be guaranteed is actually the fundamental problem of statistical learning theory*" (Hofmann, 1999). Additionally, Ding indicated that dimensionality reduction is a central and unsolved question in LSI research (Ding, 1999) (Ding, 2000).

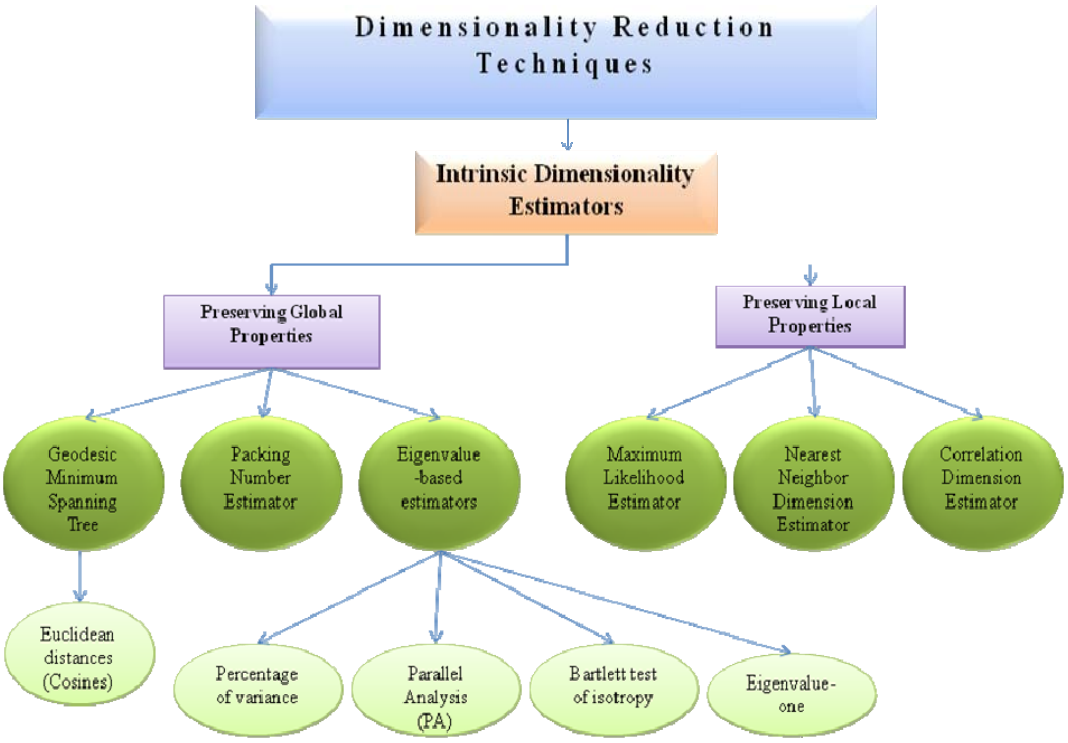


Figure 4: Taxonomy of Intrinsic dimensionality reduction techniques (*Maaten et. al. 2007*)

Table 6: Summary of published works in dimensionality reduction.

Technique / Author	Principal component analysis	Maximum likelihood factor analysis	Independent component analysis	Random projections	Self Organizing Maps	Multidimensional scaling	Neural networks	Genetic and evolutionary algorithms	Bartlett's test of isotropy	Parallel Analysis and APA	Eigenvalue-one criterion	Percent of Variance	Packing Number Estimator	Geodesic Minimum Spanning Tree
Jackson , 1991	X													
Jolliffe,1986	X													
HyvÄarinen,1999			X											
Ritter et. al.,1989				X	X									
Karhunen et. al. 1998.	X													
Kaski,1998				X										
Cox et. al. 2001						X								
Mardia, et. al., 1995						X								
Carreira-Perpina, 1997							X							
Raymer et al. 2000								X						
Fukunaga, et. al.,1971											X			
Jobson,1991									X					
Manning et al., 1999	X										X			
Borg, I. and Groenen, P. ,1997						X								
Muknahallipatna et. al, 1996							X							
Raymer et. al ,2000								X						
Kohonen,2001					X									
Jolliffe, 2002												X		
Efron, 2003										X				
Levina, et. al., 2004		X												
Burges, 2004	X	X				X								
Maaten et. al. 2007		X	X			X							X	X

2.9 Information Retrieval Systems Performance Evaluation

This research will highlight the importance of matrix dimensionality estimation technique, which will lead to the best retrieval performance. Previous research in information retrieval performance evaluation relies on a set of performance measures called Cranfield type of IR performance evaluation. Cooper (1973) stated that the goal of information retrieval evaluation is to study the performance of systems and trying to quantify their benefits. Cooper writes, "*An ideal evaluation methodology must somehow measure the ultimate worth of a retrieval system to its users in terms of an appropriate unit of utility*" (Cooper, 1973). Cranfield type of IR performance evaluation consists of a collection of experiments conducted by Cleverdon on test collections shown in table 7 (Cleverdon and Mills, 1963). Cranfield techniques are considered the most important performance evaluation techniques in IR (Salton et al., 1983), (Baeza-Yates and Ribeiro-Neto, 1999). Research on information retrieval performance evaluation includes three main components (Baeza-Yates and Ribeiro-Neto, 1999):

1) Collection of documents

2) A number of queries and

3) Group of relevance statements based on subject matter experts judgments

Table 7 : Cranfield information retrieval test collections, (Baeza-Yates and Ribeiro-Neto, 1999)

Test Collection	***Subject Matter***	***Abbreviation***
Medline	Medicine	MED
Cranfield	Aeronautics	CRAN
Communications of the ACM	Computer Science	CACM
Cystic Fibrosis (full text version) **Institute of Scientific Information**	Cystic Fibrosis (medicine) Information Science	CF FULL CISI
Cystic Fibrosis	Cystic Fibrosis (medicine)	CF

Each test collection contains a collection of documents that have been grouped by their subject of study, a collection of queries; which are statements of information needs generated by subject matter experts in the field and finally relevance judgments which are a list of all documents relevant to each query set by the panel of experts and reviewers. Statistics for each test collection in Cranfield evaluation is shown in table 8.

Cranfield type of IR performance evaluation concentrates on relevancy, since relevancy relates to the system ability to deliver related information to differentiate between relevant and non relevant documents (Harter and Hert, 1997), this problem of relevancy have been of great importance in many studies (Saracevie, 1975), (Sperber and Wilson, 1995), (Harter and Hert, 1997).

Table 8: TREC information retrieval test collections
(Baeza-Yates and Ribeiro-Neto, 1999)

Test Collection	***Abbreviation***	***# of Doc.***	***# of Terms***	***# of Queries***
Medline	MED	1033	5831	30
Cranfield	CRAN	1400	4612	225
Communications of the ACM	CACM	3200	4867	64
Cystic Fibrosis (full text version)	CF FULL	379	9549	100
Cystic Fibrosis	CF	1239	5116	100
Cystic Fibrosis Institute of Scientific Information	CISI	1460	5615	112

Cleverdon, considers relevance a function and states that for a given query q_i and a given corpus D consists of n documents d_j , $j = 1....n$, there exists a function $R(q_i, d_j)$ such that

$R(q_i, d_j) = 1$ if document j is relevant to q_i, and $R(q_i, d_j) = 0$ otherwise. Cleverdon relevancy assumption was criticized for being inaccurate and adds many contradictions and problems, since relevancy is subjective rather than objective decision. Relevancy depends on the idea and the search context, so that users can decide if a given document is relevant to their information needs. What constitutes relevant information may change over time, because we acquire more data and learn new information (Schamber, 1994). Despite its shortcomings, research done by Salton et al. (1968) and Voorhees, (1998), has demonstrated that Cleverdon objective relevance function does yield useful results for information retrieval research, since objective relevance judgments provide strong information about the benefits of one IR system over another.

Despite the relatively small size of CRANFIELD test collections shown in table 8 compared to web databases, these test collections have informed the most significant theoretical research of dimensionality reduction in IR (Ding,2000),(Hofmann, 2001), they were useful for analysis due to their variety and diversity. Since these test collections span a large area due to corpus size, domain of topics, and document representation. In general Cranfield test collections have become standard in the IR literature (Baeza-Yates and Ribeiro-Neto, 1999).

Two measures are commonly used to evaluate IR systems, Precision, which is the proportion of relevant to non-relevant documents in the retrieved documents, and Recall, which is the proportion or relevant documents in the retrieved collection to the total number of relevant documents (Van Rijsbergen, 1979) as shown in Equations 2.9.1 and 2.9.2

(2.9.1) $$PRECISION = \frac{relevant}{total\_retrieved} = \frac{|REL \cap RET|}{RET}$$

(2.9.2) $$RECALL = \frac{relevant}{total\_relevant} = \frac{|REL \cap RET|}{REL}$$

In Equation 2.9.1 and 2.9.2, REL is the set of relevant documents and RET is the set of retrieved documents. We describe $PR_{0.5}$ to be the ratio of relevant documents to the total number of documents retrieved when 50% of the relevant documents for query *q* have been retrieved. Notice that we can get a 100% recall by retrieving every document which means more cognitive burden on the user side to view many non-relevant documents in order to find relevant document matching. In general, performance is described by reporting the observed precision at a variety of recall levels (Van Rijsbergen, 1979). For example, consider the document relevancy rankings created as an example and shown in Table 9. This data set contains 40 documents. Of those 40, 20 have been judged relevant to a given query to system (A) and system (B). Table 10 lists the ordered output of these two retrieval systems.

Table 9: Example of documents relevancy ranking

System	***Document Relevancy (R:Relevant, N:Non-Relevant)***
System (A)	RNRNNRRNRNRNRNRNRNRNRNRNRNRNRNNRNRNRNRNRNR
System (B)	RNNNRRNNRRRRNNRRNNRRNRNRRRNNRRRNRNNNNNNR

Table 10: Example of documents ranking precision and recall

Recall	***System (A) Precision***	***System (B) Precision***
0.1	0.67	0.40
0.25	0.56	0.50
0.5	0.53	0.50
0.75	0.50	0.50
1	0.50	0.50

According to table 9 we calculate the precision $PR_{0.1}$ for each system by calculating the precision value for that point in the relevance ranking where ten percent (i.e. 2 relevant

documents out of 20 have been retrieved for system A and system B). We calculate precision on the ranking system A (0.1) = RNR. Thus $P_{0.1}(A) = 2/3=0.667$. On the other hand, calculating precision for the ranking in system B (0.1) = RNNNR yields $PR_{0.1}(B) = 2/5=0.4$. Thus we can say that, at the 10% recall level, system (A) yields better precision than system (B).

Usually when we evaluate a given system, we create a precision/recall curve for each point as the average precision at recall level r across each of the n queries. We plot the average precision at each recall level r by Equation 2.9.3. Precision versus recall curves for the data given in table 10 are shown in Figure 5.

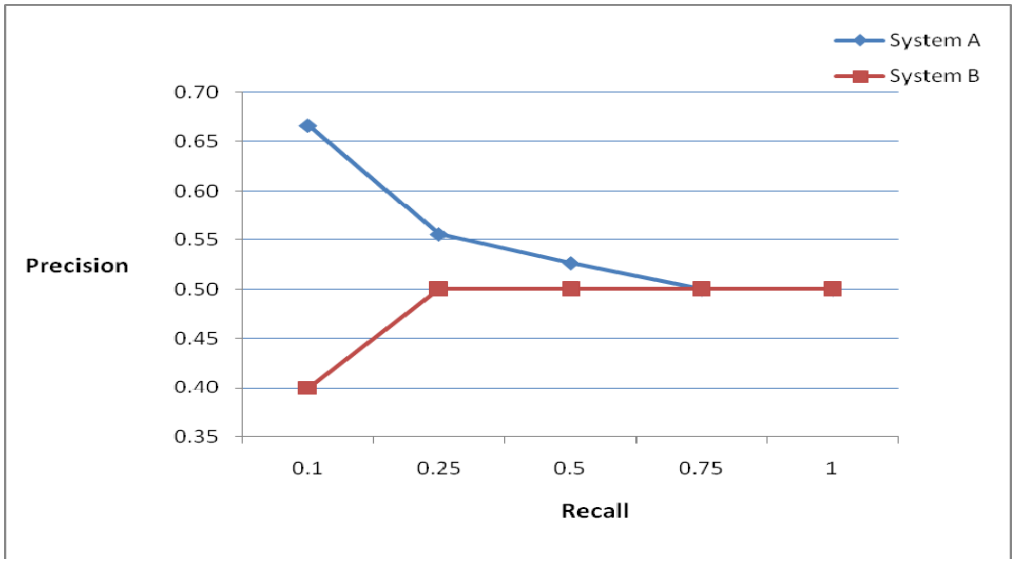


Figure 5: Precision versus recall curves for data in Table 10

(2.9.3)
$$\overline{P}_r = \sum_{i=1}^{n} \frac{P_{r,i}}{n}$$

$$r_j, j \in \{0.0,0.1,0.2,.....,1.0\}$$

As we can see from Figure 9, system (A) performs better in regards to precision than system (B) for recall levels {0.1, 0.25, and 0.5}.

We can find the interpolated precision for a given level of recall as shown in Equation 2.9.4, where the interpolated precision at the j^{th} recall level is the maximum precision at any recall level between the (j^{th}) and ($j^{th}+1$) levels (Baeza-Yates and Ribeiro-Neto, 1999).

(2.9.4) $$Interpolated\_P_r = Max\_P_r(r_j \le r \le r_{j+1})$$

Equation 2.9.5 calculates the average precision across several levels of recall, where $\overline{P}_i$, is the overall-queries average precision at recall level i , and r is the number of recall levels observed. Losee (2000) mentioned that average precision tends to provide a less biased method for information retrieval performance than other precision techniques.

(2.9.5) $$Avgerage\_P_r = \frac{\sum_{i=1}^{r} \overline{P}_i}{r}$$

Another commonly used performance measure for information retrieval evaluation is the harmonic mean of precision and recall for the j^{th} document in the ranked list of n documents which is given by Equation 2.9.6

(2.9.6) $$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

In Equation 2.9.6 $r(j)$ is the recall level for the j^{th} ranked document and $P(j)$ is the precision for the j^{th} ranked document, thus F(j) increase toward 1 when most documents are relevant and F(j)=0 until we retrieve a relevant document or F(j)=1 if all j^{th} documents are relevant. Optimal (F) is the maximum value of F found in a given system (Shaw et al., 1997) (Baeza-Yates and Ribeiro-Neto, 1999)

One of the most important information retrieval performance evaluation measures is the Average Search Length (ASL). ASL, define the expected position of a relevant document in the ranked results of an information retrieval model as shown in Equation 2.9.7 (Losee, 1998), (Losee, 2000).

(2.9.7)
$$ASL_A = \frac{\sum \mathrm{Re}l\_Position}{total\_\mathrm{Re}l}$$

If we want to calculate ASL for the example given in Table 8, then we calculate ASL by summing the position of each relevant document in each ranking and divide them by the number of relevant documents as shown in Table 11.

Table 11: Example of documents ranking average search length (ASL)

System	ASL (Document)
System (A)	$ASL_A = \frac{1+3+6+7+9+.....+38+40}{20} = \frac{407}{20} = 20.35$
System (B)	$ASL_B = \frac{1+5+6+9+10+...+33+40}{20} = \frac{384}{20} = 19.2$

As indicated in table 11, $ASL_A = 20.35$ for system A while $ASL_B = 19.2$ for system B, thus we conclude that System (B) arrange relevant documents closer to the front of the ranked list than what we get from system (A) arrangement.

The measure of Relative Relevance (RR) (Borlund & Ingwersen, 1998) is an additional performance measure which can be used to measure document relevancy for the search result in comparison to the actual document relevancy given by subject matter experts (Borlund & Ingwersen, 1998; Borlund, 2000a); relative relevancy measure equation is shown below in Equation 2.9.8.

(2.9.8) $$RR(\mathrm{Re}l_1, \mathrm{Re}l_2) = \frac{\sum(\mathrm{Re}l_1\, \mathrm{Re}l_2)}{(\sum \mathrm{Re}l_1)^{1/2} * (\sum \mathrm{Re}l_2)^{1/2}}$$

Basically, Relative Relevance (RR) measure is used in the evaluation of IR systems where more types of subjective relevance may be applied such as the well evaluated document collections provided by TREC conference. (Saracevic, 1996) (Borlund & Ingwersen, 1998) (Cosijn & Ingwersen , 2000). For example, Medline test collection has a collection of queries and identified relevant documents according to several subject matter experts, based on this we compare our dimensionally reduced IR system results for all query to get the relative relevance measure.

The RR measure evaluates the degree of agreement between results of relevance and vector cosines. Results of relevance for each query (Rel_1, Rel_2) may represent the dimensionally reduced system output, where Rel_1 represents documents ranked relevance for a specific query using a specific dimensionality reduction technique and Rel_2 represents SME's subjective relevance for each query.

The RR measure provides a more comprehensive understanding of the properties of the relevance performance of several retrieval engines, in comparison to well known relevance properties of each query being searched. The RR measure propose a solution to close the gap between subjective and objective relevance, this will reflect the effect of different dimensionality reduction techniques on the relevancy measure and its overall impact on user's cognitive load. The data in Table 12 is an example to demonstrate the implementation of the relative relevance measure for a collection of five documents.

Table 12: Example to demonstrate RR measure implementation

	Rel 1(System Relevance)	***Rel 2 (SME Relevance)***
Document 1	0.95	0.85
Document 2	0.75	0.65
Document 3	0.7	0.63
Document 4	0.64	0.4
Document 5	0.55	0.2

$$RR(\mathrm{Re}l_1, \mathrm{Re}l_2) = \frac{\sum(\mathrm{Re}l_1\, \mathrm{Re}l_2)}{(\sum \mathrm{Re}l_1)^{1/2} * (\sum \mathrm{Re}l_2)^{1/2}}$$

$$RR(\mathrm{Re}l_1, \mathrm{Re}l_2) = \frac{2.102}{(2.67)^{1/2} * (1.7419)^{1/2}} = 0.975$$

In general we can find that Cranfield information retrieval performance evaluation measures, discussed in this section, provide strong comparative evidence of whether an information retrieval system provides better performance than other systems. Since my research will include the implementation of the truncated singular value decomposition and an evaluation of different matrix reduction techniques, then Cranfield performance evaluation measures will be of much importance and guidance to this research. A summary of the published works that consider performance evaluation and dimensionality reduction in information retrieval systems is shown in Table 13.

Singular value decomposition arranges the set of documents as a vector. The task is to sort all the documents that are relevant to the user query to the beginning of the vector, and sort the non-relevant documents to the end of the vector. The question here is how much

down the ranked list will users need to consider to find all relevant documents to their search queries?

Information retrieval performance is measured by comparison to other systems. That is, the retrieval performance of a system is evaluated on a given set of documents, queries, and relevance judgments. Effectiveness of an information retrieval system is related to the relevancy of retrieved results. Relevancy, from a human perspective can be identified as a combination of the following:

- ***Subjective:*** Depends on specific user's judgment.
- ***Situational:*** Relates to user's needs.
- ***Cognitive:*** Depends on human perception and behavior.
- ***Dynamic:*** Changes over time.

This research is going to test on human labeled document collections (e.g. Medline, CRAN, and CISI) which have the following properties:

- Start with a collection of documents and a set of queries.
- Have one or more human expert to label relevant documents for each query.
- Typically assumes either one of two relevance judgments (Relevant or Non-Relevant).
- Requires considerable human effort for large document collections.

Response time is a very important factor in evaluating the usefulness of any information retrieval system. Response times of less than one second are often specified as a usability requirement. Response times are of great importance to evaluate user satisfaction in studying the interaction between computer systems and human users. Thus, an assessment of query times is a very important performance measure for an information retrieval system.

In order to do this we study the following process that a typical user will follow for query construction:

- ***User effort:*** user effort in formulating queries and screening output.
- ***Response time:*** Time interval between receipt of user's queries and presentation of system responses as shown in Figure 6.
- ***Form of presentation:*** Effect of query search output format on the user's ability to utilize the retrieved documents.
- ***Collection coverage:*** Degree to which relevant items are included in document corpus.

A typical query can retrieve hundreds to thousands of results. Results relevancy ranking is therefore a very important measure in minimizing the time spent by an individual searching for specific information thus reducing user's cognitive load. Average search length (ASL) measure defines the expected position of a relevant document in the ranked results of an information retrieval model (Losee, 1998) (Losee, 2000). The Average Search Length (ASL) measure does reflect how far users have to look in the results till they retrieve relevant documents, the less the value of ASL the better the search engine since more relevant documents will be returned in the beginning of the results which reflect a lower cognitive load on the user side and less time to be spent in filtering the results.

Standard test collections contain a set of standard documents, queries and a list of relevant documents for each query. Since this research will experiment the effect of different dimensionality reduction techniques using standard test collections, we would be most interested in measuring our system response time, which is an important performance measure that spans the time interval between receipt of a user query (in our case standard

user query) and presentation of responses. There are many mechanisms for reducing search time; our objective should be to find an acceptable trade-off between query search response time and relevancy of returned results. Human factors research indicated the need for response times faster than one second (Nielsen, 1997) (Squire et. al.1999).

Research results concerning response times in interface design is given by Nielsen (1993):

- Response time of 0.1 second is about the limit for having the user feel that the system is reacting instantaneously.
- Response time of 1.0 second is about the limit for the user's flow of thought to stay uninterrupted.
- Response time of 10 seconds is about the limit for keeping the user's attention focused on the dialogue.

In general, response time of the constructed system using different dimensionality estimation techniques will be recorded and analyzed in order to capture the effect of various dimensionality reduction techniques on retrieval performance. We would like to find the average processing time for a user query.

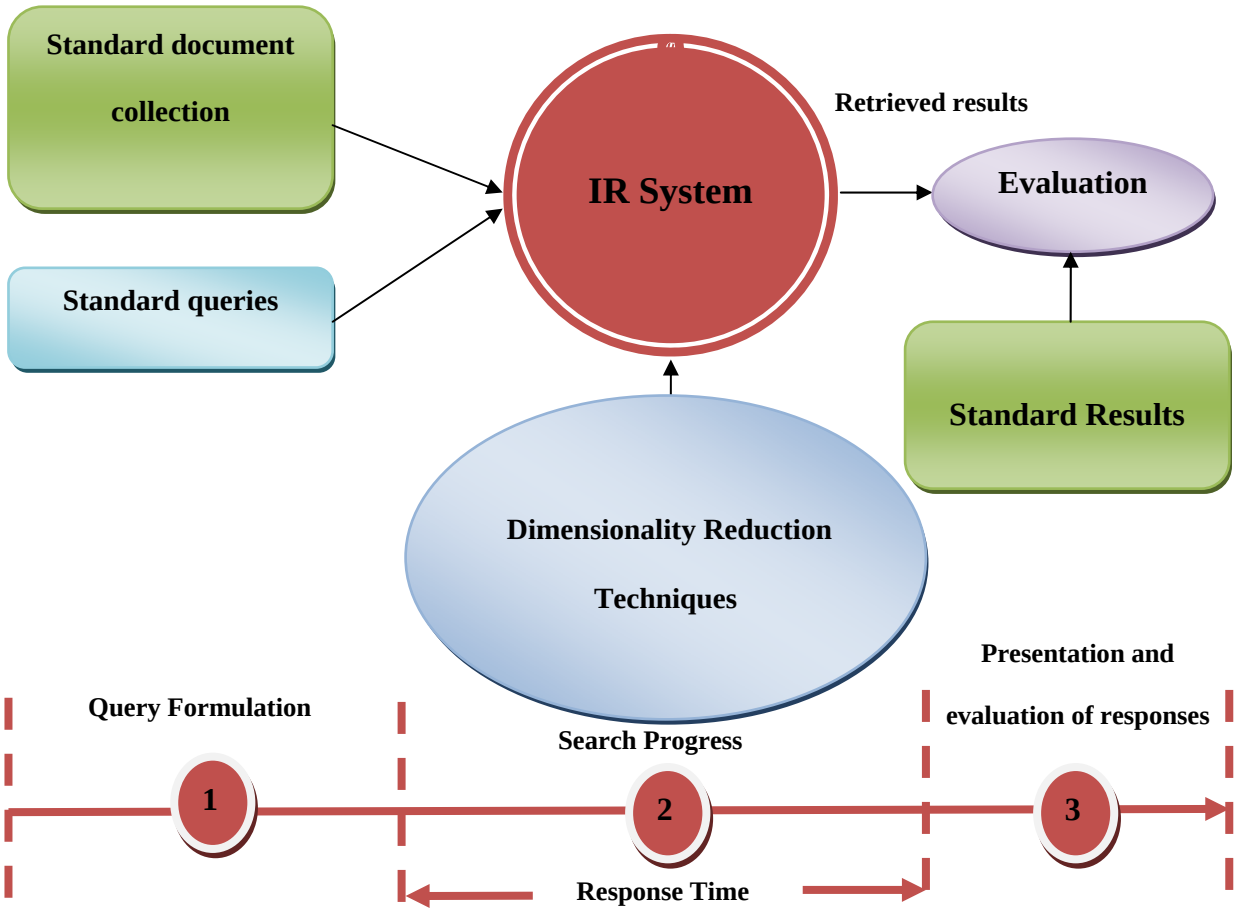


Figure 6: Dimensionally Reduced IR system response time measure

Table 13: Summary of the published works that consider dimensionality reduction in information retrieval systems

	Author	Area Researched and contribution made
Information Retrieval Systems and Intrinsic Dimensionality Estimators Research	Guttman, 1954 Fukunaga, et. al.,1971	Eigenvalue-one criterion (Kaiser-Guttman) for identifying significant principal components.
	Cleverdon and Mills, 1963	Collection of experiments conducted by Cleverdon for IR performance evaluation.
	Horn, 1965	Introduction of Parallel Analysis (PA).
	Sparck Jones, 1972	Analysis of a term's distribution across documents.
	Forsythe, et al., 1977 Golub, et al., 1989	Introduction of singular value decomposition.
	Van Rijsbergen,1977 Church et al.,1990	Observe co-occurrence in data from online corpora.
	Wittgenstein,1953 Rosch,1975 Rosch, et al.,1976	Psychological research finds a high degree of similarity among psychometric analyses of individual linguistic association.
	Salton et al., 1975 Salton et al., 1983 Salton et al., 1988	Introduction of Salton's vector space model (VSM).
	Dillon, et al., 1984 Jackson, 1993 Rencher, 1995	Introduction of the percentage of variance technique.
	Deerwester et al., 1990	Introduction of LSI.
	Jobson,1991 Anderson, 1984	Introduction of Bartlett's test of isotropy as an Eigen value based estimator for intrinsic dimensionality.
	Rencher,1995	Indicated the importance of inter-variable correlation among data collections.
	Neter, et al., 1996	Studies on LSI statistical modeling.
	Landauer et al., 1997 Landauer et al., 1998 Landauer, 2002	Studies on LSI performance.
	Ding, 1999 Ding, 2000	Study the effect of dimensionality reduction and the risk on inaccurate models.
	Manning et al., 1999	Experiments on the region of optimality with regard to parameterizing k in LSI models.
	Baeza-Yates and Ribeiro-Neto, 1999	Experiments on Cranfield type of IR performance evaluation.
	Dumais, 1993	Dumais experiments on the selection of the number of parameterizing factors.
	Story,1996	Provided a detailed discussion of the relation between information retrieval and linear regression.

Continue - Table 13: Summary of the published works that consider dimensionality reduction in information retrieval systems

	Author	Area Researched and contribution made
Information Retrieval Systems and Intrinsic Dimensionality Estimators Research	Borg, I. and Groenen, P. ,1997	Introduced Multidimensional Scaling theory and applications in dimensionality reduction
	Muknahallipatna et. al, 1996	Proposed dimension reduction in neural network training.
	Chowdhury, 1999	Use of the cosine coefficient as a measure of similarity between document and query vectors.
	Hofmann, 2000 Papadimitriou et al., 1998	Proposed the probabilistic LSA (PLSA).
	Brin, S. and Page, L. 1998	Introduction of Page Rank and Google search engine
	Kolda et al., 1997 Kolda et al., 1998 Kolda et al. 2000	Introduction of semi discrete matrix decomposition to help reduce the huge storage requirements of SVD.
	Raymer et. al ,2000	Dimensionality reduction using genetic algorithms
	Losee, 2000	Provides average precision and average search length as a less biased methods for information retrieval performance than other previously mentioned precision estimation techniques.
	Kohonen,2001	Introduced Self-organizing maps for dimensionality reduction.
	Newby, 2001 Huurnink, 2005	Research on term-based information retrieval and the side effects of undue cognitive burden placed upon end-users.
	Hofmann 2001	Criticized LSI normality assumption which is introduced by least-square method.
	Mihail et al.,2002	Research holds that term count data tend to follow a Zipf-like distribution.
	Jolliffe, 2002	Implementation of Eigenvalue-one and the percent of variance techniques for dimensionality estimation in various software packages.
	Efron, 2003	Researched Eigenvalue based estimators for dimensionality reduction and introduced Amended parallel analysis.
	Levina, et. al., 2004	Introduction of the maximum likelihood estimator for intrinsic dimensionality.
	Burges, 2004	Analyzed several methods for feature selection and dimensional reduction by dividing the methods into projective methods and manifold on which the data lies.
	Maaten et. al. 2007	Comparative study of various dimensionality reduction techniques.

2.10 The Effect of Retrieval Performance on Users Cognitive Load

Cognitive load theory (Sweller, 1988; 1994) is the instructional theory that describes human learning structures in terms of information processing. This includes long term memory, which stores all of our skills and knowledge permanently and working memory, which continues to perform and supervise tasks associated with consciousness. Information may only be stored in long term memory after first being processed by working memory. Working memory limitations will impede overall due to its effect on both capacity and duration.

Cognitive load have been used with little understanding of Cognitive Load Theory. Cognitive Load Theory (CLT) has been introduced and developed by educational psychologists such as Sweller (1988; 1994). IR can be viewed as a problem solving process with which users try to solve their information search problem by query formulation (Kuhlthau, Spink, and Cool, 1992). Cooper (1998) indicated that Cognitive Load Theory can be used to describe structures of learning and patterns of thinking.

Copper (1998) stated that "*cognitive load theory focuses on the role of working memory in the learning process*". The fundamental principles of cognitive load theory rely on the following (Back and Oppenheim, 2001):

- Working memory is limited.
- Long term memory is essentially unlimited.
- The process of learning requires working memory to be actively engaged in the comprehension (and processing) of instructional material to encode to-be-learned information into long term memory.

- If the resources of working memory are exceeded then learning will be ineffective.

Cognitive load theory has been used with IR research in reference to Human Computer Interaction issues. The only IR study that has tried to include the concepts of cognitive load theory was performed by Hu, Ma, and Chau (1999). Based on their research they examined the effectiveness of designs using wither a graphical or list-based concepts that best supported the communication of an object's relevance. Cognitive load was used in research as a measure of information processing effort a user must provide to take notice of the visual stimuli in an interface and understand its influence (Hu, Ma, and Chau, 1999). In previous studies it was assumed that users would prefer an interface design that requires low cognitive load in general and at the same time, can result in high user satisfaction with more relevant results. Various reporting methods were used to match individual users assessments of the cognitive load associated with a particular interface. However, this research will try to demonstrate, that the concept of cognitive load associated with information retrieval systems can be extended beyond interface design to include the effect of dimensionality reduction when considering multiple performance measures in enhancing query search results.

Although Back and Oppenheim (2001), Kuhlthau (1993) mentioned that there are many components for cognitive load in information retrieval they discussed three main components:

- ***Retrieval Performance:*** Indicates that cognitive load increases as the number of relevant documents identified by the system increase. This research will concentrate on evaluating the effect of selecting proper dimensionality reduction parameters on enhancing overall search performance.

- ***User's knowledge of the information need***: Cognitive load reduces as more domain knowledge is gained.

- ***User's overall level of doubt:*** Level of uncertainty associated with the search process. Cognitive load reduces as users become aware that their information need can be addressed.

Back and Oppenheim (2001) referred to information uncertainty as a cognitive stage which causes anxiety and lack of confidence that leads to cognitive load. Uncertainty due to a lack of understanding or miss-interpreting the meaning initiates the process of information seeking (Kuhlthau, 1993).

Since this research will concentrate on evaluating the effect of information retrieval on the search performance, we will involve the evaluation of the performance of different systems by selecting the dimensions found by several dimensionality reduction techniques. Cognitive load is related to the effectiveness of an IR system since it can be measured in terms of how long it takes for a user to reach relevant information or reach the conclusion that no relevant information exists. A search query can return thousands of results. Thus document relevancy is a very important measure in minimizing the time spent by the user searching for specific information and will help reducing overall cognitive load during information search process. Average search length measure defines the expected position of a relevant document in the ranked results of an information retrieval model (Losee, 1998) (Losee, 2000). We calculate ASL by summing the position of each relevant document in each ranking and dividing by the number of relevant documents. Additionally average search length measure does reflect how far the user have to look in the results till he get

relevant documents, the less the value of ASL the better the search engine since more relevant documents will be returned in the beginning of the results which will be reflected by lower cognitive load on the user side. Using document relevancy will reduce the problems caused by information overload through avoiding large number of documents returned to the user. It is recommended to limit the size of information returned in order to prevent distracting the user from answering his search question or requiring extensive filtering. This implies a technical reduction of the quantity of information by dimensionality reduction to minimize the noise or distraction introduced by large documents collection.

2.11 Evidence of Research Gap

Information retrieval can be viewed as a problem solving process with which users try to solve their information search problem by query formulation (Kuhlthau, Spink, and Cool, 1992). Latent semantic analysis reflects human knowledge since its results are similar to those of humans on standard vocabulary and subject matter expert tests. Additionally latent semantic analysis simulate human word sorting, category judgments and estimates content coherence, learnability of information by individual student users, and the quality and quantity of knowledge included in an essay (Landauer, Foltz, and Laham, 1998). LSI can be used as a reliable method for the representation of word meaning that produces measures of word-word, word-document and word-concept relations that are similar to much human cognitive aspects involving association and representational similarity.

As discussed earlier, Cognitive Load is related to the effectiveness of an IR system and can be measured in terms of how long it takes for a user to reach appropriate and relevant information, or discover that no relevant information exists. A typical query can

retrieve thousands of results. Document relevancy ranking is therefore a very important measure in minimizing the time spent by an individual searching for specific information thus reducing cognitive load during search process. Intrinsic cognitive load is related to the difficulty of tasks, while extraneous cognitive load is related to the presentation of tasks (Cooper, 1998). Modifying task presentation to a lower level of extraneous cognitive load will minimize problem solving effort if the resulting total cognitive load falls to a level within the range of cognitive resources.

As the size and dimensionality of data increases, query performance diminishes and this is usually reflected and measured by the average system precision. Literature review of research in dimensionality reduction indicted that no one to date has researched the effect of different dimensionality reduction methods on user's cognitive overload, measured through multiple IR performance measures.

Researchers have found that dimensionality reduction provides a better solution to IR problems, which results in faster response times, with reasonable accuracy and precision. A good dimensionality reduction technique has the capability of reducing the data into a lower-dimensional model, while maintaining the properties of the original data. Therefore it is desirable to find which technique provides better estimates for data dimensionality in order to improve user's cognitive performance, especially in dense information environments such as the World Wide Web, while preserving important information from the original data collection. One common way to reduce data dimensionality is to project the data onto a lower-dimensional subspace. Previous research done on information retrieval systems using LSI has generally found improvements in search results, however still there is a lack of research which detail and evaluate the effect of dimensionality reduction on

reducing user's cognitive load. The main problem is that there is no consensus about the most effective method for estimating the best number of dimensions in LSI and there is a need for more research to be conducted on evaluating the effect of dimensionality on a set of performance measures.

This research is concerned with the parameterization of *k*, the number of retained dimensions during the implementation of singular value decomposition. Additionally, this research will test and compare the effect of different dimensionality reduction techniques on information retrieval systems overall performance using a set of performance measures.

Due to the importance of dimensionality reduction, a number of new techniques for dimensionality reduction have been proposed recently in image processing. A systematic empirical listing of a large number of dimensionality reduction techniques has been presented in Maaten et. al. (2007), such techniques have not been researched for the implementation in information retrieval systems to improve query search results. Document relevancy as a performance measure is expected to reduce the problems caused by information overload through avoiding large number of documents returned to the user. It is recommended to limit the size of information returned in order to prevent distracting the user from answering selected search question or requiring extensive filtering. This implies a technical reduction of the quantity of information by dimensionality reduction to minimize noise or the distraction introduced by large data collections.

As stated above, the context of this research is the selection of the number of dimensions retained using dimensionality estimation algorithms that will improve overall search performance. Although latent semantic indexing has seen many successes, there is still much unknown on the effect of dimensionality reduction on enhancing search

performance. Intuition suggests that using reduction techniques to select the proper dimension would be so important to achieve better search results. However, applying this to large data collections is complex and therefore we should perform theoretical investigations and thorough examination of the results of practical dimensionality reduction algorithms on selected document test collections.

Based on this discussion we arrive at the following problem statement. Under what conditions can a specific dimensionality reduction algorithm improve query search while reducing user's cognitive load?

To answer this question, different methods have to be studied in detail in order to study their characteristics and effects on search performance. As a guideline to this research the following research questions have been formulated:

(1) Theoretical properties of dimensionality reduction methods,

(2) Characteristics of efficient implementations in term of results relevancy and other performance measures which impact search performance,

(3) The best dimensionality reduction technique that will result in better overall system performance and reduced cognitive load?

This research will seek a better structure of the data collection to uncover concepts associations, which are hidden as semantic properties. Because of the complexity of this type of research, theoretical research alone will not be able to answer the problem statement. We need to find out whether suggested dimensionality reduction methods are of practical use. Therefore, they have to be implemented into standard test collections and the properties

of those dimensionality estimation methods have to be investigated with respect to various performance measures.

If efficient implementations of the search methods are possible, different techniques have to be tested in realistic experimental conditions because the final answer to our problem statement depends on whether and when these techniques work effectively in practice.

This research is going to contribute in identifying the best dimensionality estimation method which will reduce user's overall cognitive load by enhancing retrieval performance in terms of relevancy and better concept matching, additionally, novel dimensionality estimation techniques will be introduced and tested against other methods. Results will help answering several questions such as: what is the best dimensionality reduction technique that will result in better overall system performance?

This research will look for enhanced dimensionality reduction techniques that will improve matrix dimensionality estimation and enhancing search results in terms of increasing relative relevance, precision and recall while reducing average search length and query processing time; this will reduce the time it takes the user to find specific information and will reduce users level of uncertainty and doubt associated with the search process since the cognitive load will be reduced as users becomes more confident that their information need can be addressed.

CHAPTER THREE: IR MULTI-CRITERIA DECISION ANALYSIS

A decision is a choice made such that selected alternatives are the best among other possible candidates. The decision process is not always easy. Most of the time, there are many criteria's to base the judgment on and no alternative can be found to outrank all others under each performance criteria. Decision makers also have to prioritize and weight the relative importance of selected criteria in order to achieve agreement on selected alternatives. In IR systems we encounter the problem of making a decision to select one alternative or system over another based on selected performance measures.

Previous research in IR performance evaluation considered precision and recall as the primary, and sometimes sole, performance measures to decide on overall system performance, in doing so, they ignored the impact of relative relevance, average search length and time on overall system performance. This chapter will discuss the effect of multi-criteria decision analysis (MCDA) on information retrieval performance and will introduce a novel method based on MCDA to enhance query search and overall performance ranking.

3.1 Multi-criteria Decision Analysis (MCDA)

Decision analysis is a group of systematic procedures for studying and analyzing complex decision problems (Malczewski, 1997). Multi-criteria decision analysis (MCDA) methods have been designed to select and rank alternatives according to a set of criteria's (Lootsma, 1999). Malczewski(1999) divided multi-criteria decision analysis into three steps: 1) Design phase, where decision rules and preferences are specified and alternatives are

considered, 2) Choice phase, where sensitivity analysis is used to gain better insight about the problem, 3) Intelligence phase where decision matrix is studied and criteria's are evaluated. Multi-criteria decision analysis process is shown in Figure 7.

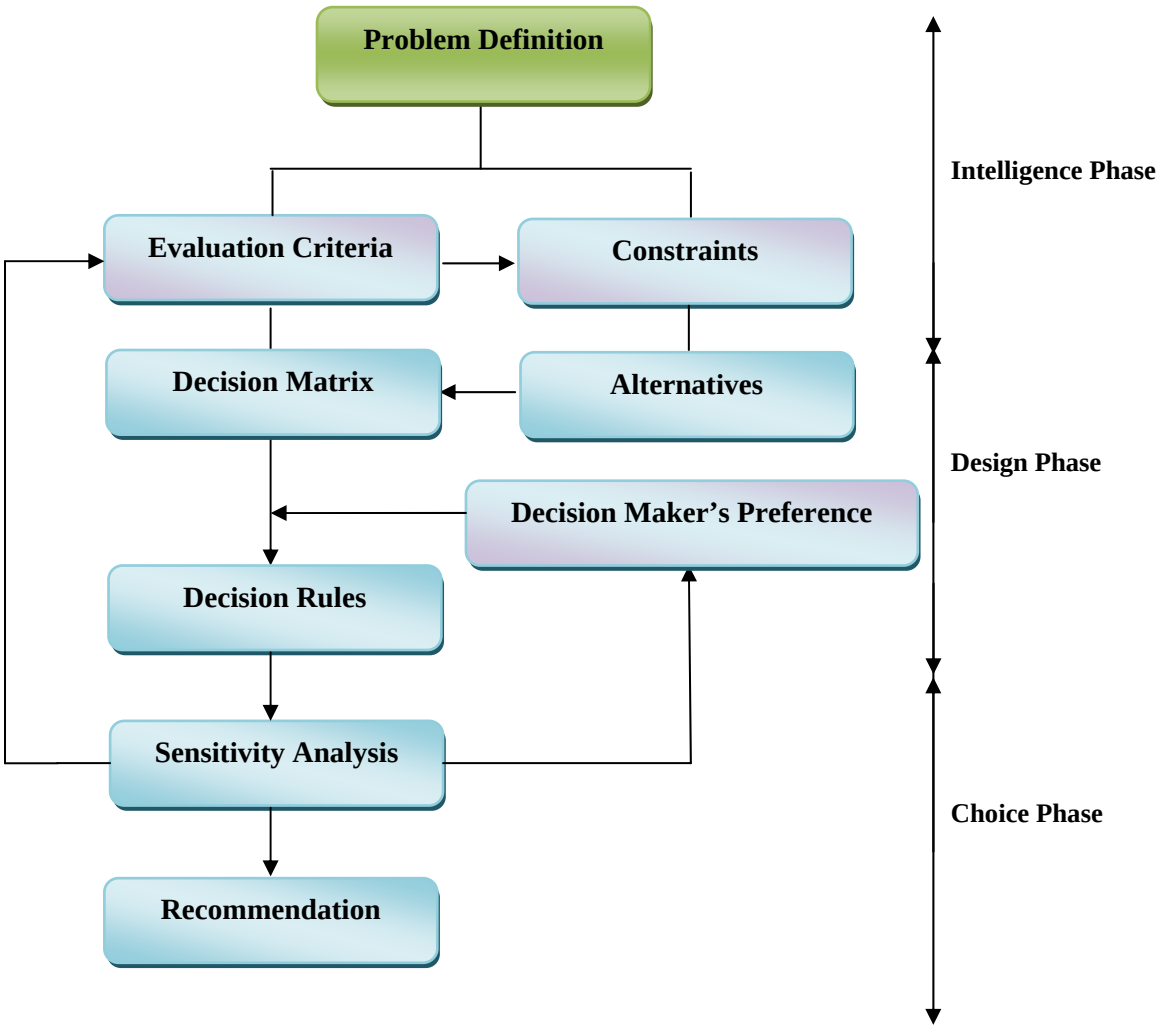


Figure 7: Framework for MCDA process used in GIS system (Malczewski, 1999)

MCDA techniques can be used to rank alternatives, list a number of options for evaluation, identify most preferred alternative, or to differentiate between acceptable and unacceptable selections (Dodgson, 2000) (Malczewski, 1997). MCDA techniques can be classified as either Multi-Objective decision making (MODM) or Multi-Attribute decision making (MADM). The difference between MODM and MADM is based on the evaluation criteria,

which is a general term and includes both attributes and objectives, for which an attribute is a measurable quantity whose value reflects the degree to which a particular objective is achieved. Objectives are derived from a set of attributes as a statement about the desired state of the system. (Malczewski, 1997) (Malczewski, 1999). Tables 16 and 17 summarize the most often used MODM and MADM methods. Various methods used in estimating weights are discussed below.

3.2 Criterion Weights Assignment

Information retrieval performance measures have weights that reflect the values assigned to performance measures and indicate their relative importance compared to other measures under consideration. Weights assignment to performance measures provides an indication to the different degrees of importance for each performance measure. There are four different techniques for assigning weights: Ranking, Rating, Pairwise Comparison and Trade of Analysis, Table 15provide a summary of weighing methods.

3.2.1 Ranking Methods

Ranking method is a simple method for evaluating the importance of multiple performance measures based on ranking each criterion in the order of decision maker's preferences. Ranking method disadvantages are: lack of theoretical foundation and inappropriateness when used with larger number of performance measures (Malczewski, 1999).

3.2.2 Rating Methods

Rating method asks the decision maker to estimate weights on the basis of a predetermined scale (Malczewski, 1999). This method rely on allocating points ranging from zero to one hundred, where zero indicates that the criterion can be ignored, and a hundred represents the

situation where only one criterion needs to be considered. Or a score of one hundred is assigned to the highly important criterion and proportionally smaller weights are given to criteria lower in order. Disadvantages are: lack of theoretical foundation and also the assigned weights might be difficult to justify.

3.2.3 Pairwise Comparison Method

This method implements pairwise comparisons as input and produce relative weights as output, advantages of this method is that only two criteria's have to be considered at a time. Pairwise comparison disadvantage is that if you have many criteria's, the amount of pairwise comparisons that should be made will be very large. Pairwise comparison involves three steps (Malczewski, 1999) (Saaty, 1980):

(1) Create pairwise comparison matrix using a scale with values ranging from (1 to 9) as shown in Table 14.

(2) Computation of the weights in three steps:

   I. Calculating the summation of the values in each column of the matrix,

   II. Dividing each element in the matrix by its column total to get the normalized pairwise comparison matrix,

   III. Computation of the average of the elements in each row of the normalized matrix.

(3) Estimation of the consistency ratio to determine if the comparisons are consistent. This can be done through the following steps:

   I. Calculation of the weighted sum vector by multiplying the weight for the criterion times the column of the original pairwise comparison matrix, then sum these values over the rows,

II. Find the consistency vector by dividing the weighted sum vector by the criterion weights determined previously,

III. Compute the average value, lambda (λ), of the consistency vector and Consistency Index (CI), this average provides a measure of departure from consistency and has the following formula (Malczewski, 1999):

$$CI = (\lambda - n)/(n-1)$$

IV. Calculation of the Consistency Ratio (CR) which is defined as follows:

$$CR = CI / RI$$

Where: *RI* is the random index and depends on the number of elements being compared. If CR<0.10, the ratio indicates a reasonable level of consistency in the pairwise comparison, however, if CR ≥ 0.10, the values of the ratio indicates inconsistent judgments (Malczewski, 1999).

Table 14: Pairwise Comparison Scale (Saaty, 1980)

Intensity of Importance	Definition
1	Equal importance
2	Equal to moderately importance
3	Moderate importance
4	Moderate to strong importance
5	Strong importance
6	Strong to very strong importance
7	Very strong importance
8	Very to extremely strong importance
9	Extreme importance

3.2.4 Trade-Off Analysis Method

Trade-off analysis involves a comparison between two alternatives with respect to two criteria's at a time and assessment of which alternative is preferred. A unique set of weights will be defined that will allow all preferred alternatives in the trade-offs to get the same overall value. A disadvantage of this method is that the decision maker is presumed to follow axioms to make final judgments (Malczewski, 1997).

Table 15: Methods used in estimating weights (Malczewski, 1999)

Method	***Ranking***	***Rating***	***Pairwise Comparison***	***Trade-off Analysis***
No. Judgments	*n*	*n*	*n(n-1)/2*	*<n*
Response scale	*Ordinal*	*Interval*	*Ratio*	*Interval*
Hierarchical	*Possible*	*Possible*	*Yes*	*Yes*
Underlying Theory	*None*	*None*	*Statistical/ Heuristic*	*Axiomatic/ Deductive*
Ease of use	*Very easy*	*Very easy*	*Easy*	*Difficult*
Trustworthiness	*Low*	*High*	*High*	*Medium*
Precision	*Approximations*	*Not precise*	*Quite precise*	*Quite precise*
Software Availability	*Spreadsheets*	*Spreadsheets*	*Expert Choice*	*Logical Decision*

3.3 Analytical Hierarchy Process (AHP)

Analytical hierarchy process is a decision support technique developed by Saaty(1980) for analyzing and supporting decisions for situations with multiple competing objectives and alternatives. AHP is based on three main steps:

1. ***Decomposition:*** decision problem is decomposed into simpler decision problems to form a decision hierarchy (Erkut and Moran, 1991). The hierarchy decreases from the general goal to more specific levels until a level of attributes are reached. Hierarchical structure includes four levels: goal, objectives, attributes and alternatives,
2. ***Comparative judgment:*** using pairwise comparisons to reduce the complexity of decision making problem,
3. ***Synthesis of priorities:*** combine the relative weights of the levels obtained in the above step by multiplications of the matrices of relative weights at each level of the hierarchy. The matrix is squared and the row sums are calculated and normalized for each row in the comparison matrix. This sequence is continued when the difference between the normalized weights of the iterations become smaller than a determined value (Saaty, 1990).

3.3.1 Evaluation of IR Systems Overall Performance Using AHP

Wang and Forgionne presented a decision-theoretic approach based on AHP for evaluating IR systems from a user perspective and reported its workability and proofed AHP suitability to IR evaluation with promising results. (Wang and Forgionne, 2005).

Godwin (2000) used AHP to model and study information technology (IT) outsourcing decisions. Results indicated that AHP can be used effectively to analyze IT decisions and provides a computer based group decision environment needed to capture experts' opinions on several criteria's. The sensitivity analysis of AHP is important in that it creates real-time,

interactive, graphical display of the ranking of the options as the decision makers compare between different possibilities.

Kawasaki and Sunahara, acheived improved response time of distributed multemedia retrieval network through the use of integrated AHP into query routing system (Kawasaki and Sunahara,2000). Based on the results from previous studies for using AHP in various IR problems, AHP enhanced systems performance and improved decision and alternatives ranking.

3.4 Multi-Criteria Weighted Model to Estimate Intrinsic Dimensionality

In estimating term document matrix intrinsic dimensionality we encounter the problem of making a decision to select a cutoff value (k). Many alternatives and techniques exist and all claim increased performance for a selected measure. This involves a decision making problem to select an alternative over the other based on selected performance measures. Inspired by the work done in the field of Multi-criteria Decision Analysis, this research propose a novel method to estimate matrix intrinsic dimensionality based on using a multi-criteria model for weighted performance measures.

In the proposed multi-criteria weighted model we calculate the sum of weighted values of k which gave us best performance using all possible dimensions. In order to achieve best performance we seek maximizing precision, recall and relative relevance while minimizing query processing time and average search length. Thus we multiply the value of k which gave maximum precision by the weight of precision as a performance measure assigned by SME's, doing the same for all other performance measures and taking the summation, as shown in Equation 3.3.1.1, is expected to give a better estimate for intrinsic dimensionality

that accounts for system overall performance. In Equation 3.3.1.1, calculations of $(k_{Pr}, k_{Rc}, k_{RR}, k_{ASL}, k_t)$ is based on the experimental results for selected test collections using various dimensionality estimation techniques. Thus $k_{Max\,\mathrm{Pr}}$, is the value of *k* that resulted in the maximum overall precision using the selected dimensionality estimation technique.

(3.3.1.1)

$$k_{Weighted} = \sum[(W_{\mathrm{Pr}} \times k_{Max\,\mathrm{Pr}}) + (W_{Rc} \times k_{MaxRc}) + (W_{RR} \times k_{MaxRR}) + + (W_{ASL} \times k_{MinASL}) + (W_t \times k_{Min.t})]$$

Where:

$k_{Pr} = k_{\text{Max Precision}}$, $k_{Rc} = k_{\text{Max Recall}}$, $k_{RR} = k_{\text{Max Relative Relevance}}$,
$k_{ASL} = k_{\text{Min Avg. Search Length}}$, $k_t = k_{\text{Min Query Processing Time}}$

W_{Pr} *: Priority of precision performance measure*

W_{Rc} *: Priority of recall performance measure*

W_{RR} *: Priority of relative relevance performance measure*

W_{ASL} *: Priority of average search length performance measure*

W_t *: Priority of query processing time*

Although decision-making theories have existed for a long time, the application of decision science especially AHP into information retrieval systems to evaluate overall performance is a new contribution to the field of information retrieval. The weighted multi-criteria model for leveraging the effect and weight of multiple performance measures is anticipated to provide a better estimate of intrinsic dimensionality based on accounting for overall system performance.

Table 16: Summary of the most often used MODM methods (Malczewski, 1999)

MODM Method	***Input***	***Output***	***Types of Decision***	***DM Interactor***	***Assumptions***
Value/ Utility model	Value/Utility Functions, Weights	Best alternative	Individual DM, deterministic, probabilistic	Moderate/ high	Very restrictive
Goal Programming	Aspiration Levels, Priorities, weights	Best Alternative	Individual DM, deterministic, fuzzy	High	Very restrictive
Interactive Programming	Aspiration reservation	Satisfying alternative	Individual DM, deterministic, fuzzy	Moderate increases with problem size	Moderately restrictive
Compromise Programming	Ideal point, Weight	Compromise alternative, cardinal ranking	Individual and group DMs, probabilistic, fuzzy	Moderate	Moderately restrictive
Data Envelopment Analysis	Set of evaluation inputs and outputs	Cardinal ranking	Individual and group DMs, deterministic, probabilistic, fuzzy	Low	Moderately restrictive

Table 17: Summary of the most often used MADM methods (Malczewski, 1999)

MADM Method	Input	Output	Types of Decision
Scoring (SAW)	Attribute scores, weights	Ordinal ranking	Individual DM, deterministic
Multi-attribute value	Value functions, weights	Cardinal ranking	Individual and group DMs, deterministic, fuzzy
Multi-attribute utility	Utility functions, weights	Cardinal ranking	Individual and group DMs, probabilistic , fuzzy
Analytic hierarchy process	Attribute scores, pairwise comparisons	cardinal ranking (ratio scale)	Individual and group DMs, deterministic, probabilistic , fuzzy
Ideal point	Attribute scores, weights, ideal point	Cardinal ranking	Individual and group DMs, deterministic, probabilistic , fuzzy
Concordance	Attribute scores, weights	Partial pr ordinal ranking	Individual and group DMs, deterministic, probabilistic , fuzzy
Ordered weighted averaging	Fuzzy attribute, weights, order weights	Cardinal or ordinal ranking	Individual and group DMs, fuzzy

CHAPTER FOUR: AVERAGE STANDARD ESTIMATOR (ASE)

This chapter will introduce the Average Standard Estimator (ASE), a novel method for estimating data intrinsic dimensionality based on singular value decomposition. ASE estimates the level of significance for singular values resulted from the truncated singular value decomposition (TSVD). Truncated singular value decomposition proceeds by including only those significant singular values according to ASE and excluding those with low significance. In doing so we include the analysis of term independence discussed in Chapter 2, since singular values reflect terms dependence, a lower value of ASE reflects more terms independence as will be shown in this chapter.

4.1 The Method of Average Standard Estimator in IR Systems

The basic assumption behind latent semantic analysis and truncated singular value decomposition is that term correlation in information retrieval reduces searchers cognitive burden. LSI was created to address the gap between information spaces and cognitive spaces so as to improve data representation to accommodate for the error of term independence (Landauer et al., 1997), (Landauer et al., 1998), (Foltz et al., 1998),(Gardenfors, 2000),(Landauer, 2002),

Several researchers referred to the deficiency of current information retrieval methods, in which, the words searchers use in their queries are not the same as those by which the information they seek has been indexed, this will result in relatively poor search performance. As discussed in chapter 2, latent semantic indexing use a low rank approximation of the original data matrix by adopting the use of singular value

decomposition (SVD), a least-squares matrix factorization method from linear algebra (Golub, et al., 1989), (Forsythe, et al., 1977), (Berry et al., 1994), (Strang, 1998).

Wong (1987) generalized vector space model (GVSM) improved retrieval results by assuming terms non-orthogonality and interdependence. This assumption of terms interdependence is proven to be true by other researchers (Manning et al., 1999) (Oakes, 1998) (Cooper, 1988) (Cooper, 1991).

SVD is used to derive a least-squares approximation of matrix A, as shown in Equation 2.4.2, where all term-document similarities are approximated by the results of this model with the reduced dimension (Deerwester, 1990). In Equation 4.1 and 4.2, matrix $\sum$ is an $r \times r$ diagonal matrix, with the diagonal elements $sv_{11} \geq sv_{22} \geq sv_{33} \geq ..... \geq sv_{rr} \geq 0$ called the singular values (Deerwester, 1990) (Berry et al., 1994) (Hastie et al., 2001). The matrix of singular values $\sum$ acts as a reference when selecting singular values to retain during dimensionality reduction.

(4.1) $$A = T_{[m \times r]} \sum_{[r \times r]} D_{[r \times n]}$$

(4.2) $$\sum_{(rxr)} = \begin{pmatrix} sv_{11} & 0 & ... & 0 \\ 0 & sv_{22} & ... & 0 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ 0 & 0 & ... & sv_{rr} \end{pmatrix}$$

To estimate intrinsic dimensionality of the sparse matrix, we have to differentiate between large and small singular values. Selecting proper singular values involves deriving a suitable method for judging their significance based on their magnitude.

ASE is based on the concept of terms correlation represented by singular values in SVD, thus if terms in the document collections are independent then there will be no improvement by dimensionality reduction. However, as discussed earlier, terms dependency is proofed to be true in previous research. Noticing that calculated singular values decrease in a magnitude of different rates, average standard estimator (ASE) is concerned in the cutoff point, where the calculated singular value magnitude decrease in a rate less than the average rate. The proposed method, overcome shortcomings of previous methods by selecting a cutoff value based on analyzing all singular value rate of decrease in magnitude, then ASE select those values which satisfy this condition shown in Equation 4.3.

$$\text{Average Decrease in Magnitude} = \frac{\sum_{m=1}^{r-1} sv_{(m+1)} - sv_{(m)}}{r-1} \tag{4.3}$$

In order to account for random noise distracters in the data, we add a multiplier (n) of singular values standard deviation to the cutoff average estimator. This is helpful since it leads to a dynamic estimation of k. Thus, for document collections with relatively small size of indexed terms, selecting a higher standard deviation multiplier (e.g. 1.5 or 2) reflects the need to account for less variability in the data; this will include the effect of small singular values and prevent ignoring important relationships. While for larger data collections, with respect to indexed terms, adding a lower value of standard deviations multiplier to the average estimator (e.g. 0 or 0.5) will result in a decline of those factors corresponding to

relatively small singular values which contain essentially random noise distracters, this approach align with Ding (1999, 2000) and Story (1996) research recommendations to improve search performance by accounting for the effect of random noise distracters. The average standard estimator model is shown in Equation 4.4.

$$ASE = \frac{\sum_{m=1}^{r-1} sv_{(m+1)} - sv_{(m)}}{r-1} + (n)S.D \tag{4.4}$$

ASE estimates the number of dimensions retained in the truncated singular value decomposition shown in Equation 4.5 by including only those singular values in the data set which are larger or equal to the cutoff point estimation based on Equation 4.4. In Equation 4.5, T_k contains the fist k columns of T estimated by ASE and $\sum_k$ contains the first k rows and columns of $\sum$ estimated by ASE, and D_k contains the first k columns of D estimated by ASE.

$$\hat{A}_{k(ASE)} = T_{k(ASE)} \sum_{k(ASE)} D_{k(ASE)} \tag{4.5}$$

Additionally the effect of selected value for the standard deviation component in ASE will be studied for three test collections, recommendations will be suggested based on document characteristics and overall IR system performance results.

While previous research in dimensionality reduction underestimates document collections intrinsic dimensionality. ASE technique is useful since it applies a practical rationale to estimate intrinsic dimensionality. ASE method remedy the underestimation problem of intrinsic dimensionality in previous approaches by accounting for standard deviation as an important factor to accommodate for variability in document collection characteristics and in regard to the number of documents and indexed terms. ASE assumes that variables in the

document collection with deep relations have sufficient correlation and that only those relationships with high singular values are significant and should be maintained. Based on this discussion and preliminary data analysis shown in the next section, ASE is expected to improve matrix intrinsic dimensionality estimation by including the effect of both singular values magnitude of decrease and random noise distracters.

4.2 Example of Dimensionality Estimation Using ASE

This section will discuss an example of using the average standard estimator to estimate data sets intrinsic dimensionality. In this example we tested ASE on MEDLINE document collection for the first 15 queries, and compared the results obtained under ASE with those obtained under Kaiser-Guttman technique and dimensionality estimation based on scree plot using only the first (10) most relevant documents returned by the dimensionally reduced system for each query. Using scree plot to estimate MEDLINE intrinsic dimensionality (k), we find that intrinsic dimensionality was estimated approximately at ($k_{SP} = 203$) as indicated in Figures 8 and 9.

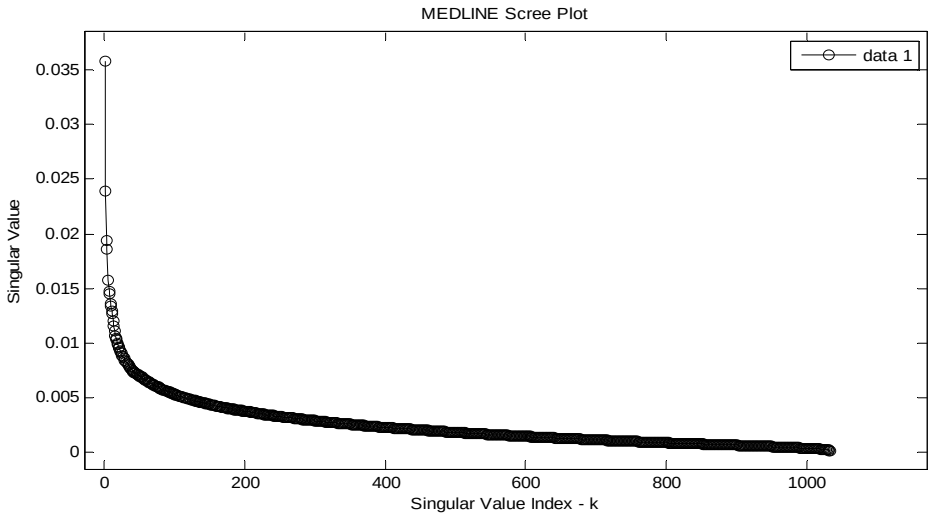


Figure 8: Scree Plot for MEDLINE test

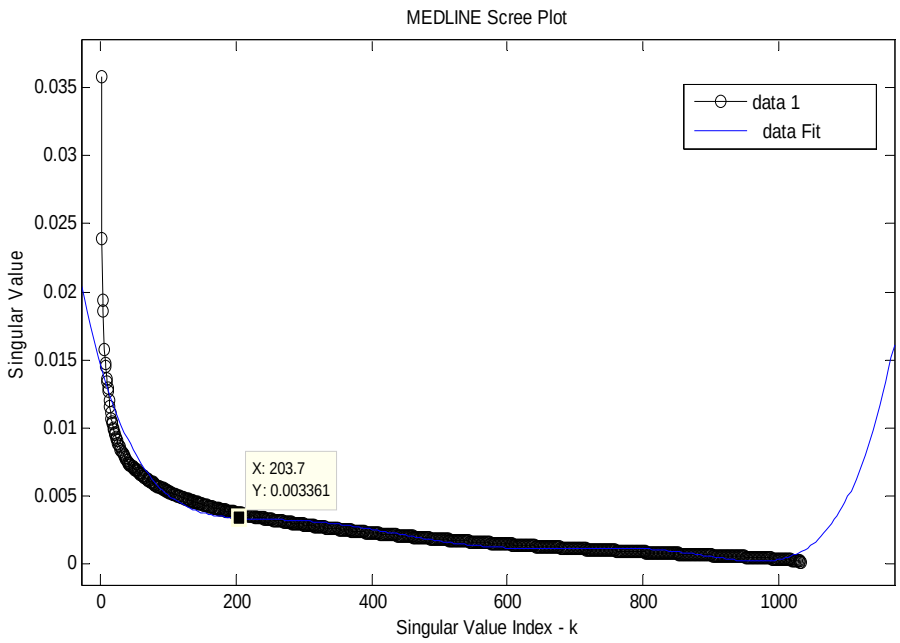


Figure 9: Scree Plot with data fitting for MEDLINE test collection

Kaiser-Guttman technique retains all factors whose corresponding singular values are greater than the average of all the singular values (Guttman, 1954), using this technique, MEDLINE intrinsic dimensionality was estimated at ($k_{KG} = 358$).

Average standard estimator technique (ASE) estimates MEDLINE data intrinsic dimensionality at ($k_{ASE} = 182$) using (n=1.5) for the standard deviation multiplier. MEDLINE document collection has a relatively small size of indexed terms compared to other document collections, thus selecting a relatively high standard deviation reflects the need to account for more variability in the data; this will include the effect of smaller singular values and prevent ignoring important relationships.

Results for the three intrinsic dimensionality estimators with various performance measures is included in Appendix (B) and summarized in Table 18 below. Appendix (C) includes sample Matlab code used to generate results for this example.

Table 18: Summary of performance measures using ASE compared to two other intrinsic dimensionality estimators for 15 queries in MEDLINE test collection

Method	***K***	***Average Precision***	***Recall***	***ASL***	***Relative Relevance***	***Average processing time/query***
Kaiser-Guttman	358	0.6800	0.3835	1.9044	0.9034	3.7565
ASE (n=1.5)	182	0.7133	0.3979	1.8782	1.0542	1.7233
Scree plot	203	0.6933	0.3929	1.8380	1.0388	1.9427

Based on the results of this example, we conclude that ASE achieved better estimation of matrix intrinsic dimensionality with regard to average precision, recall and improvement in query processing time; however, these results are not conclusive since they were based on small testing scale. This example highlights the need to develop a model to assess and evaluate overall dimensionality estimation performance with regard to various evaluation measures. Additionally, we would like to evaluate collections overall performance under various values of standard deviations to find the relation between various document collections characteristics and selected ASE parameters.

CHAPTER FIVE: PROPOSED METHODOLOGY

Research discussed in this document is concerned with the parameterization of k, the number of retained dimensions during the implementation of truncated singular value decomposition. Analysis is aimed at discovering a better and effective means for selecting k in unsupervised environments while maintaining a reasonable query response time for information retrieval systems. This research will try to give answers to the following question: Can we achieve better search results in terms of relative relevance, precision and recall, while reducing search time and average search length through the use of the weighted multi-objective set of performance measures to achieve an improved estimate of matrix intrinsic dimensionality? To be able to achieve a better estimate of the matrix dimensionality, there is a need to study a number of document collections and evaluate each test collection using a number of performance measures.

Since there is no agreement on which performance measure is the best mean to assess retrieval performance, this research suggested a new technique to evaluate search overall performance based on a multi-criteria weighted model. We start by estimating k using various dimensionality estimation techniques in addition to the multi-weighted model and the novel dimensionality estimation technique based on the Average Standard Estimator (ASE) using average distances between consecutive singular values and n standard deviations as a cut-off value to estimate k. After getting various estimates of k for each document collection, result will be processed in the truncated singular value decomposition using various performance measures including the multi-criteria weighted model and compare the results using the analytical hierarchy processing (AHP).

5.1 Information Retrieval Test Collections

When comparing intrinsic dimensionality estimation methods, Kolda et al. (2000) used three standard document collections as indicated in Table 19. MEDLINE (MED) is a collection of 1033 medical abstracts from the Medlars collection. CISI is a collection of 1460 information science abstracts. CRANFIELD (CRAN) is a collection of 1398 aerodynamics abstracts from the Cranfield collection. Each test collection comes with a collection of documents, a collection of queries, and the correct answers to each query is a list of relevant documents. Those three test collections have been selected because they cover major types to test collections with different characteristics.

Those test collections were recommended by TREC because they have been evaluated and studied by experts and used in previous theoretical research in IR systems as standard document collections. Selected test collections, MED, CISI, CRAN, were also recommended because they have been used in Ding's theoretical work on dimensionality reduction for IR (Ding, 1999) (Ding, 2000). Thus using these documents collection for this study allows comparison with previous results obtained under other studies for similar kind of problems.

When evaluating a query, we get an ordered list of documents with the position of those documents in the ordered list reflects relevancy to the search query. For each query, we compute the recall and precision values in addition to relative relevance, ASL and query search time. Selecting document collections with different numbers of documents will ensure capturing the relationship between terms, documents and concepts among various collections.

Table 19: Characteristics of selected document test collections (Source: Kolda et al., 1997)

Characteristics	***MEDLINE***	***CRANFIELD***	***CISI***
Number of Documents:	1033	1399	1460
Number of Queries:	30	225	112
Number of (Indexing) Terms:	5526	4598	5574
Avg. No. of Terms/Document:	48	57	46
Avg. No. of Documents/Term:	9	17	12
% Nonzero Entries in Matrix:	0.87	1.24	0.82
Storage for Matrix (MB):	0.4	0.6	0.5
Avg. No of Terms/Query:	10	9	7
Avg. No Relevant/Query:	23	8	50

We compare various IR systems by looking at various IR performance measures such as average precision, recall, ASL, relative relevance and response time, some of which are standard measures used by the information retrieval community (Harman 1995), (Kolda, 1997). The first and second rows of Table 19 reflect the number of documents and the number of queries in each test collection. Third row reflects the number of indexing terms in each test collection. Selected document collections have different characteristics with regard to the number of documents and the size of their indexed terms. The number of indexing terms is the number of selected terms used to represent each document after processing documents and removing stop-words. Rest of Table 19 describes other document collection characteristics, such as the average number of terms per document and average number of relevant documents per query as studied by SME's. By selecting document collections with varying numbers of documents per term, we would like to ensure that different relationships between terms, documents and concepts among the various collections will be captured.

Those three collections, with various sizes regarding the number of documents and vocabulary were selected in order to maximize the diversity of experimental characteristics, and to ensure capturing different relationships between terms, documents and concepts. It is also of great interest to study how the value of estimated k relates to different features of a data set, and how various dimensionality estimation techniques perform on data of various features.

5.2 Information Retrieval Performance Measures

Finding IR models that enhance document retrieval performance requires observing retrieval performance in terms of various performance measures including Cranfield-based metrics discussed earlier in Chapter 2. The objective is to find an information retrieval system with better collective system performance as will be discussed in this section.

5.2.1 Cranfield Performance Measures

IR performance evaluation metrics have been selected for a number of reasons, first, average precision has become a common criteria and standard performance indicator in IR research, thus, defining performance in terms of precision and recall is preferred since it aligns this research results with the majority of previous research in the field as discussed in Chapter 2.

After estimating intrinsic dimensionality for selected test collections, MED, CISI, CRAN using various estimation techniques, this research will test these findings using a system built for this purpose and analyze different results using Cranfield information retrieval performance evaluation measures and other measures based on experimental results. Three other metrics, Relative Relevance (R.R), Average Search Length (ASL) and search time will be included since they will help assessing the validity of results, and in order to find whether

observed best performance with regard to precision, recall, ASL, relative relevance and time agrees on the matrix intrinsic dimension. Different approaches for dimensionality reduction will then be applied to each test collection and evaluation will be based on selected IR performance measures as shown in Table 20.

Table 20: IR selected performance measures

Measure	***Description***
Precision (Prec.)	Average precision at various recall levels.
Recall (Rec.)	Average recall per query.
Relative Relevance (R.R)	Ratio of IR system and expert relevance for returned document.
Average Search Length (ASL)	Location of a relevant document in the ranked output of search result.
Response time (t)	Average IR system query processing time.

According to Losee (2000), average precision tends to provide a less biased method for information retrieval performance than other previously mentioned measures shown in Equations 2.9.1 and 2.9.5. Recall measure, shown in Equation 2.9.2, is the proportion or relevant documents in the retrieved collection to the total number of relevant documents (Van Rijsbergen, 1979). This research will adapt the mathematical formulation of both average precision and recall as denoted by Losee (2000), Van Rijsbergen(1979) and Baeza-Yates and Ribeiro-Neto(1999).

Relative Relevance ($R.R$) measure, shown in Equations 2.9.7, is used in the evaluation of IR systems where more types of subjective relevance may be applied such as the well evaluated document collections provided by TREC conference. (Saracevic, 1996) (Borlund & Ingwersen, 1998) (Cosijn & Ingwersen , 2000). The R.R measure computes the degree of

agreement between two results of relevance assessments or vectors cosines. This research will adapt the mathematical formulation of relative relevance which was suggested by Cosijn & Ingwersen (2000). For example, MEDLINE test collection has a collection of queries and identified relevant documents for each query according to subject matter experts, based on this we can compare our dimensionally reduced IR system results for each query to get the relative relevancy score.

Average Search Length (ASL), define the expected position of a relevant document in the ranked results of an information retrieval model (Losee, 1998), (Losee, 2000). As discussed in Chapter 2, we calculate ASL by summing the position of each relevant document in each ranking and dividing by the number of relevant documents as shown in Equation 2.9.7. Since our goal is to find a suitable trade-off between response time and the quality of retrieval results. Response time will be used as a performance measure to determine the usefulness of information retrieval systems. In general, Cranfield information retrieval performance measures, provide strong comparative evidence of whether an information retrieval system provides better performance than other systems and since this research will include the implementation of the truncated singular value decomposition and an evaluation of different matrix reduction techniques, then Cranfield performance evaluation measures will be of much importance and guidance to accomplish such objectives.

5.2.2 Evaluation of IR Overall Performance Using Analytical Hierarchy Processing

In AHP we specify different evaluation measures and integrate them into a multi-criteria hierarchy, AHP model identifies the factors that must be measured to evaluate the effectiveness of an IR system from a decision-making perspective (Wang and Forgionne,

2005). The hierarchy in AHP isolates the specific cause of a decision outcome leading to more objective results. For many MCDM methods the effects of dissimilarities in weights produced by these methods become obvious leading to inconsistent results for problems with few alternatives. (Zanakis et al., 1998). Based on the research conducted by Wang and Forgionne (2005), Godwin (2000), Kawasaki and Sunahara (2000), AHP proofed suitability to information retrieval. Results indicated that AHP can be used effectively to analyze IT decisions and provides a computer based group decision environment needed to capture experts' opinions on several criteria. The sensitivity analysis of AHP is important in that it creates real-time, interactive, graphical display of the ranking of the options as the decision makers compare between different possibilities. Although decision-making theories have existed for a long time, the application of decision science especially AHP into information retrieval to evaluate systems overall performance is a new contribution to the field of IR systems performance evaluation.

5.3 Dimensionality Estimation Techniques

Results from literature review and current research on dimensionality estimation highlights several techniques for intrinsic dimensionality estimation which can be used to improve performance; a list of these techniques is shown below:

- Kaiser-Guttman technique.
- Singular Values estimation based on scree plot.
- Percentage of variance explained (90%).

- Average Standard Estimator (ASE): a novel dimensionality estimation technique for estimating intrinsic dimensionality based on the average distance between consecutive singular values and n standard deviations to estimate k .
- Intrinsic dimensionality estimation based on the results of the Multi-criteria weighted model developed in this research.

5.4 Methodology Outline:

Research experimental framework is illustrated in Figure 10 and summarized in the following steps:

- Converting text documents to Term-Document Matrix
- Parsing standard queries using terms indexes (Tokens) from TD matrix
- Calculate singular value decomposition of TD matrix
- Apply ASE dimensionality estimation technique to find the reduced dimension estimate "k".
- Apply all other dimensionality estimation techniques to find the reduced dimension estimator (k).
- Update calculated Singular value decomposition to include only the k highest singular values resulted from each dimensionality reduction method.
- Calculate selected performance measures for each standard query in each test collection.
- Calculate performance measures for all queries in each document collection.
- Calculate the weighted importance of each performance measure using the relative importance scale from the ranked results of subject matter experts.

- Use the Multi-criteria weighted model developed in this research to find a new estimate of k and apply the estimated k value to each document collection and compare results.
- Use Analytic Hierarchy Process (AHP) to evaluate and compare different dimensionality estimation techniques according to the results of their performance measures.

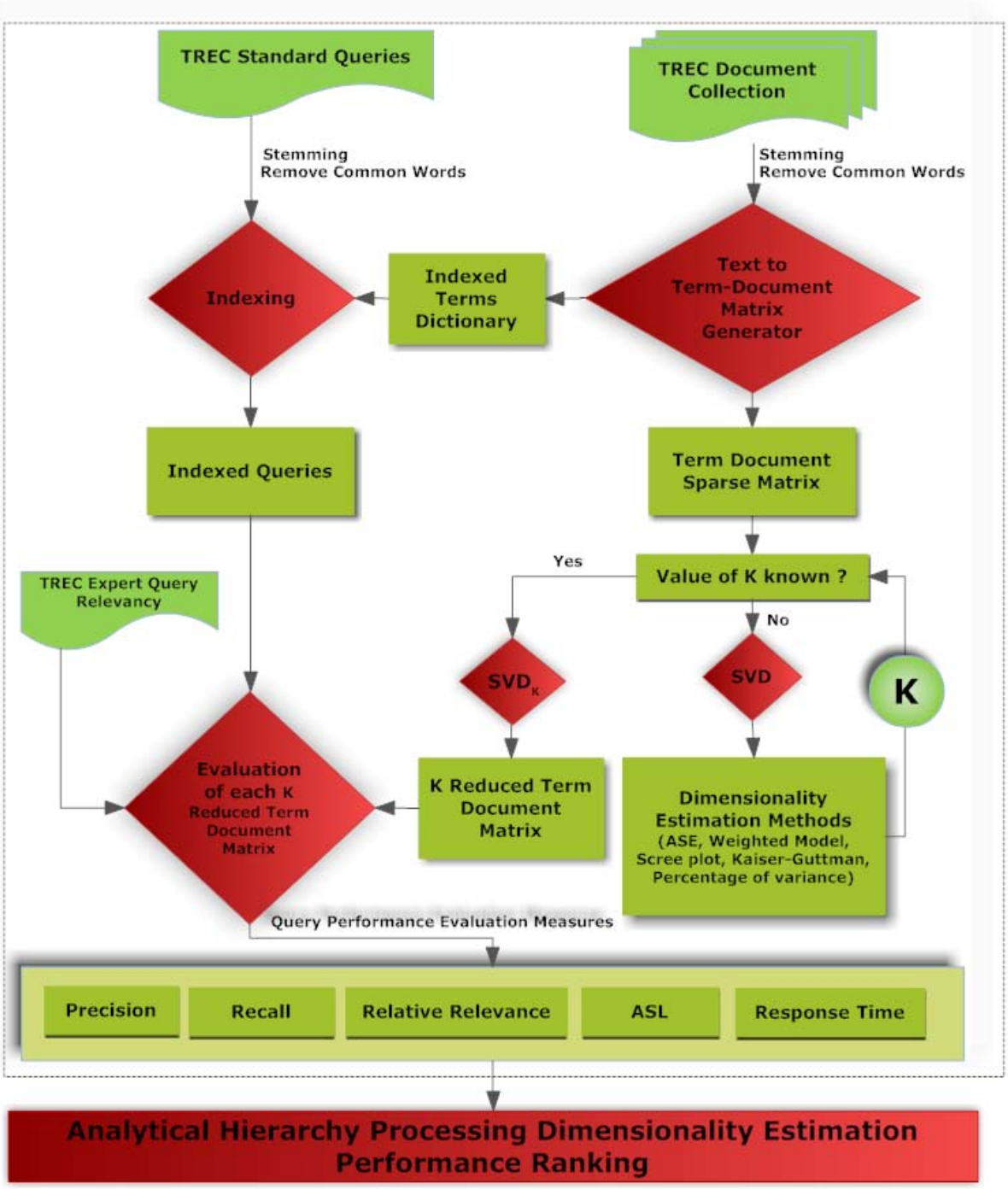


Figure 10: Framework of proposed methodology

5.5 Software and Computational Tools Used In Experimentation

This section reviews selected hardware and software choices that have been made to enable successful experimentation and computation of various performance measures. The objective is to provide guidance and proof of workability and accuracy of selected tools for future researchers interested to work in this field.

All experimentations were performed on the latest generation of computers supported with Intel Core 2 processors with CPU's clock speed of 2.20GHz, equipped with up to 4 gigabytes of physical random access memory, 350 gigabytes of disk space, and 512MB dedicated graphics card. This hardware configuration will allow the system to perform matrix indexing, singular value decomposition and truncated singular value decomposition on a small size term by document matrix in less than half an hour of CPU time.

All of the software used to calculate estimated intrinsic dimensionality, performance measures and all other calculations will be in *MATLAB R2007a Version 7.4.0*. MATLAB stands for "*Matrix Laboratory*", which is a mathematical computing software from *Math Works*.

The indexing module included in Text to Matrix Generator (TMG) will be used in MATLAB to generate term by document sparse matrices. One of the benefits of TMG is that it can be used for the construction of new and the update of existing term document matrices from text collections in the form of MATLAB sparse arrays. Choices made for term by document generation and terms indexing were based on the default recommended choices in common indexing standards in Text to Matrix Generators. Table 21 lists default indexing choices.

Table 21: Text to term document selected parameters (Zeimpekis and Gallopoulos, 2007)

Parameter	*Description*	*Selected Value*
Delimiter	The delimiter between tmg's views of documents. Possible values are 'emptyline', 'none delimiter' (treats each file as single document) or any other string.	Empty line and "I."
Stop list	Name of file containing stopwords, i.e. common words not used in indexing.	SMART , English Common Words
Min Length	Minimum term length.	3
Max Length	Maximum term length.	30
Min Local Frequency	Minimum local term frequency.	1
Max Local Frequency	Maximum local term frequency.	Inf.
Min Global Frequency	Minimum global term frequency.	1
Max Global Frequency	Maximum global term frequency.	Inf.
Local Term Weighting	Local term weighting function. values: 'Term Frequency'(TF)	TF
Global Term Weighting	Global term weighting function. Possible values: 'None','Entropy', 'Inverse Document Frequency (IDF)', 'GfIdf', 'Normal', 'Probabilistic Inverse'.	None

To perform SVD and TSVD, this study will use functions that are based on LAPACK routines, which are provided in LAPACK library. LAPACK is a library of Fortran 77 subroutines for solving many problems in numerical linear algebra and designed to be efficient on a wide range of modern high-performance computers (Anderson et al., 1999). Library download is available at (http://www.netlib.org/lapack/). The Statistics Toolbox functions and basic routines will be used to help complete the logic and construct testing

codes as shown in ASE code example in Appendix (C). Analytical Hierarchy Processing will be performed using Expert Choice Software ver. 11.5, Expert Choice software can be downloaded from:(http://www.expertchoice.com/).

Original code for the calculation of all related performance measures and dimensionality estimations techniques will be developed based on the literature review in chapter 2 and the discussion of the two novel methods in Chapters 3 and 4 (ASE and Multi-criteria weighted model). Figure 11 shows a sample MATLAB code.

```
% Open TD matrix.
data=X
% Size of TD matrix.
[n,p] = size(data);
% Center the data.
datac = data - repmat(sum(data)/n,n,1);
% Find the covariance matrix.
covm = cov(datac);
[svec,sval] = eigs(covm,p);
% find SV for the first 1032 (k<n) row and column
sval = diag(sval);  % extract the diagonal elements
% order in descending order
svec = svec(:,p:-1:1);
% Draw a plot.
figure, plot(1:length(sval),sval,'ko-')
title('MEDLINE Scree Plot')
xlabel('Singular Value Index - k')
ylabel('Singular Value')
```

Figure 11: Sample SVD MATLAB code.

5.6 Analysis of Results

Experimentation will provide complex results to be studied and analyzed. In Chapter 6, experimental results will be analyzed to rank various dimensionality estimation methods according to their overall performance for various measures. Of particular interest, the Average Standard Estimator method and Multi-criteria weighted model proposed in this research which provided improved performance in estimating data intrinsic dimensionality.

CHAPTER SIX: RESULTS

The previous chapter provided detailed outline for testing document collections using various dimensionality estimation techniques, and rank them according to different performance measures. In addition to introducing a new approach based on the multi-criteria weighted model. A novel dimensionality estimation technique was introduced based on the Average Standard Estimator (ASE), this is a new technique which have been proposed in this research for estimating data intrinsic dimensionality (k) that corresponds to the average distance between consecutive singular values and a multiplier of standard deviations. Following are the dimensionality estimation techniques tested in this research:

- Kaiser-Guttman technique;
- Intrinsic Dimensionality estimation based on scree plot;
- Percentage of variance explained (90%);
- Average Standard Estimator (ASE);
- Intrinsic dimensionality estimation based on the Multi-criteria weighted model.

Since estimating data intrinsic dimensionality though Cranfield performance measures requires much attention, this chapter starts with a general overview and comparison of document collections retrieval performance under various dimensions. Section 6.2 will include detailed analyses of strengths and weaknesses of each dimensionality estimation method. Finally section 6.3 will summarize experimental results and findings.

6.1 Overview of Experimental Outline

This section gives only a brief discussion on the methods used for data analysis after experimentation and testing. Analysis of results will involve three major parts. Studying and analyzing the singular value decomposition performance results at a range of selected dimensionalities on the three selected document collections. This will give us insight and indication about the effect of selected dimensionality on various performance measures. After estimating the intrinsic dimensionality for each document collection, this study will evaluate and analyze the effect of k value on the system overall performance and how estimated values correlates with performance analysis of k value.

The performance of various dimensionality estimation techniques, including the average standard estimator (ASE) and intrinsic dimensionality estimation will be researched based on the multi-criteria weighted technique and AHP analysis. Additionally, it is of interest to study the effect of selected value of the standard deviation multiplier in ASE on retrieval performance. Recommendations will be suggested based on overall IR system performance while analysis will involve all document collections described in Table 19. Intrinsic dimensionality will be estimated by five estimation techniques. Results will be validated through comparison between various dimensionally reduced IR systems through search results and *TREC* standard document relevancy (i.e. expert's relevancy ranking for each query). The second stage involves a comparison of the results for all dimensionality estimation techniques and study performance for the multi-criteria weighted estimation of k .

This research will help facilitate better understanding for the strengths and weaknesses of each dimensionality estimation method and analyze the effect of document collection characteristics on overall system performance. Studying the effect of documents

characteristics such as sparsity on IR systems performance will facilitate better understanding to the relation between various factors in term of matrix size, sparsity, and value of k on each performance measure.

6.2 Intrinsic Dimensionality Estimation for Document Collections

This section provides analysis of test collections to find the best representative dimension for data intrinsic dimensionality. We are interested in knowing if whether, for a given test collection, a low-rank approximation of the term-document matrix will improve model performance over a full-rank model. If matrix dimensionality reduction will improve a system's precision, recall, ASL and relevance measures, then it would be interesting to find which value of k led to the most noticed improvement. Also we are going to use multiple performance measures to find how dimensionality reduction improves retrieval performance over a full rank model. It's crucial to notice that the amount of dimensionality reduction required to find matrix intrinsic dimensionality varies across test collections.

The problem of estimating dimensionality for documents collection with different characteristics is part of current research problems in dimensionality estimation, this research will try to avoid conflicting measures by introducing the multi-criteria weighted model to reach the dimension(s) that satisfy multiple performance objectives. Table 23 summarize findings for test collections intrinsic dimensionality estimation (k_{Est}) with respect to various performance measures.

For each performance measure in Table 23 there are four statistics: the value of k that led to best performance (k_{Est}) with respect to the selected performance measure, actual value of selected performance measure observed at its respective k_{Est} , amount of dimensionality

reduction from the full rank model for selected performance measure and percentage of total variance covered when selecting k_{Est} .

6.2.1 Analytical Hierarchy Processing (AHP) Model Results

A multi-weighted performance measures model was constructed based on the results of the analytical hierarchy processing (AHP) for performance measures ranked by subject matter experts. Ranking details provided in Appendix D. AHP analysis indicates that precision overall priority in the information retrieval system as (0.128), recall overall priority (0.156), relative relevance overall priority (0.235), average search length overall priority (0.235) and processing time overall priority (0.245). AHP analysis was conducted using *Expert Choice Software v11.5*; results are summarized in Table 22 and Appendix E.

Table 22: Summary AHP results using SME's ranking

AHP Performance Measures priorities based on SME's ranking (Inconsistency=0.08)		
Processing Time Priority	0.245	
Relative Relevance Priority	0.235	
Average Search Length Priority	0.235	
Average Recall Priority	0.156	
Average Precision Priority	0.128	

Based on experimental findings for the three document collections shown in Table 23 and Equation 3.3.1.1, we calculate $k_{Weighted}$ for each documents collection as shown below:

$$k_{Weighted} = \sum[(W_{\Pr} \times k_{\Pr}) + (W_{Rc} \times k_{Rc}) + (W_{RR} \times k_{RR}) + +(W_{ASL} \times k_{ASL}) + (W_t \times k_t)]$$

Where:

$k_{Pr} = k_{\text{Max Precision}}$, $k_{Rc} = k_{\text{Max Recall}}$, $k_{RR} = k_{\text{Max Relative Relevance}}$, $k_{ASL} = k_{\text{Min Avg. Search Length}}$,
$k_t = k_{\text{Min Query Response Time}}$ where : $k_t = Min(t)[k_{Pr}, k_{Rc}, k_{RR}, k_{ASL}]$

W_{Pr} *: Priority of precision performance measure from AHP analysis.*

W_{Rc} *: Priority of recall performance measure from AHP analysis.*

W_{RR} *: Priority of relative relevance performance measure from AHP analysis.*

W_{ASL} *: Priority of average search length performance measure from AHP analysis.*

W_t *: Priority of query processing time from AHP analysis.*

$$k_{Weighted\_MEDLINE} = \sum[(0.128\times150)+(0.156\times150)+(0.235\times100)++(0.235\times90)+(0.245\times90)]$$
$$k_{Weighted\_MEDLINE} = 109$$

$$k_{Weighted\_CRANFIELD} = \sum[(0.128\times320)+(0.156\times320)+(0.235\times320)++(0.235\times100)+(0.245\times100)]$$
$$k_{Weighted\_CRANFIELD} = 214$$

$$k_{Weighted\_CISI} = \sum[(0.128\times1350)+(0.156\times1250)+(0.235\times850)++(0.235\times350)+(0.245\times350)]$$
$$k_{Weighted\_CISI} = 736$$

We find that the multi-weighted model estimates MEDLINE intrinsic dimensionality at k=109 and CRANFIELD intrinsic dimensionality at k=214 and CISI intrinsic dimensionality at k=736. In the next section those estimates, which were obtained by the multi-weighted methods, will be compared with other results obtained by various dimensionality estimation techniques.

Table 23: Summary of document collections intrinsic dimensionality estimation k_{Est} with respect to multiple performance measures.

Characteristics:	**MEDLINE**	**CRANFIELD**	**CISI**
Number of Documents	1033	1399	1460
Number of Queries	30	225	112
Number of (Indexing) Terms	5526	4598	5574
Average Number of Terms/Query	10	9	7
Average Number of Relevant Documents /Query	23	8	50
Number of Documents Returned	10	10	10
K_{Est} **(Precision),** (Percentage of total dimensionality retained)	**150** (14.5%)	**320** (22.9%)	**1350** (92.5%)
Precision at k_{Est}	0.680	0.156	0.278
Dimensionality Difference $(k_{Est} - k_{Max})$	-883	-1079	-110
Variance Captured at k_{Est} *(Precision) (%)*	26.33%	61.97%	99.2%
Average Processing (Seconds)time at k_{Est}*(Precision)*	20.88	40.67	257.54
K_{Est} **(Recall),** (Percentage of total dimensionality retained)	**150** (14.5%)	**320** (22.9%)	**1250** (85.6%)
Recall at k_{Est}	0.331	0.2187	0.1127
Dimensionality Difference $(k_{Est} - k_{Max})$	-883	-1079	-210
Variance Captured at k_{Est} *(Recall) (%)*	26.33%	61.97%	97.92%
Average Processing time (Seconds) at k_{Est} *(Recall)*	20.88	40.67	236.105
K_{Est} **(ASL),** (Percentage of total dimensionality retained)	**90** (8.7%)	**100** (7.1%)	**350** (23.9%)
ASL at k_{Est}	1.580	0.828	0.4135
Dimensionality Difference $(k_{Est} - k_{Max})$	-943	-1299	-1110
Variance Captured at k_{Est} *(ASL) (%)*	17.45%	33.10%	57.41%
Average Processing time (Seconds) at k_{Est} *(ASL)*	23.709	13.25	63.69
K_{Est} **(Relative Relevance),** (Percentage of total dimensionality retained)	**100** (9.6%)	**320** (22.9%)	**850** (58.2%)
Average Relative Relevance at k_{Est}	1.127	0.208	0.4279
Dimensionality Difference $(k_{Est} - k_{Max})$	-933	-1079	-610
Variance Captured at k_{Est} *(R.R) (%)*	19.01%	61.97%	87.78%
Average Processing time (Seconds) at k_{Est} *(R.R)*	13.805	40.67	158.61

6.2.2 Test Collections Experimental Results

Results shown in Table 23 and Figure 12 for Medline test collection indicate the variation and disagreement between various performance measures. There is a clear disagreement on the value of k_{Est} for a selected document collection. Figure 13 indicates the relationship of query processing time with the number of dimensions retained, this relationship highlight the need to retain the minimum number of dimensions that will result in the best overall model performance within a reasonable query processing time.

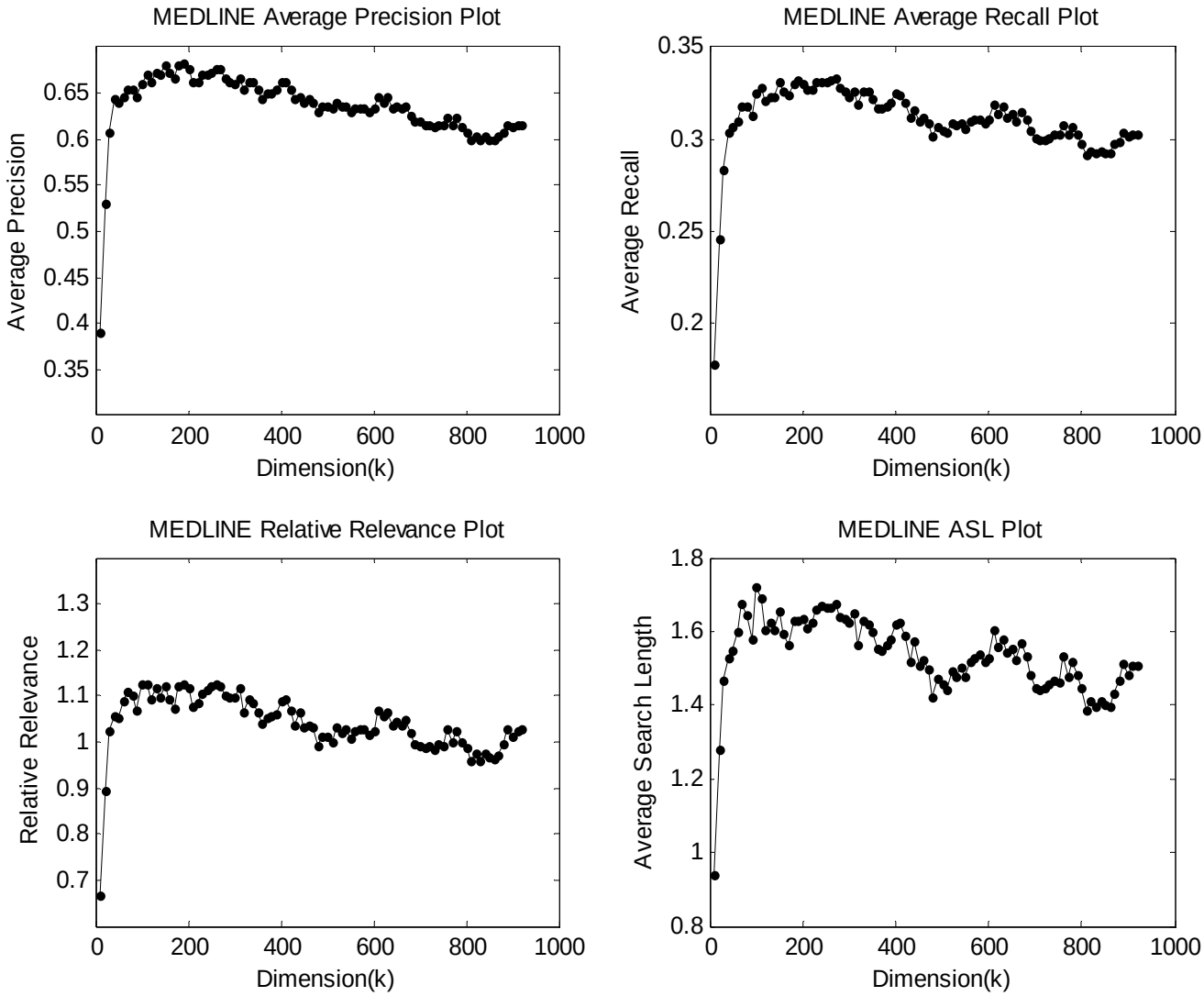


Figure 12: MEDLINE performance results

Experimental results for MEDLINE indicate that k_{Est} should be in the vicinity of 150 to 200, this will provide the highest performance with respect to average precision as shown in Figure 12. It is important to notice that performance for all measures increase as the number of dimensions retained increase up to a certain point (intrinsic dimensionality). Past this point performance starts to decrease.

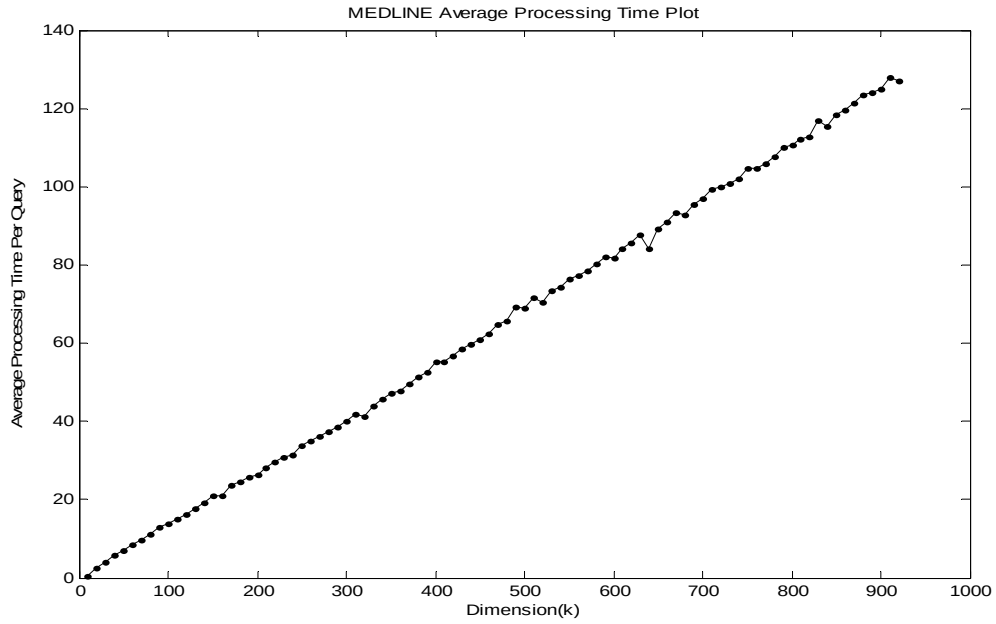


Figure 13: MEDLINE average query processing time (Seconds)

Performance measures have been studied across all possible MEDLINE matrix dimensions. It was noticed that average search length and average query processing time measures performed the best at lower dimensions, while average precision, average recall and relative relevance have close agreement on higher dimensions. This research will seek a good balance between each performance measure to achieve best overall retrieval performance.

From Table 23, using average precision and average recall performance measures in MEDLINE, it is clear that k_{Est} (Avg. Precision) was similar to k_{Est} (Avg. Recall) at k_{Est} =150, this is 14.5% of full rank model. While average search length (ASL) performance measure in MEDLINE performed best at k_{Est} (ASL) =90, this is 8.7% of full rank model. Relative relevance performance measure in MEDLINE performed the best at k_{Est} (R.R) =100, this is equivalent to 9.6% of the full rank model.

As discussed in Chapter four, the Average Standard Estimator (ASE) is concerned in the cutoff point, where the calculated singular value rate of change is less than the average rate of change. The negative effects of random noise distracters will be minimized by adding a multiplier (n) of singular values standard deviation to the cutoff point calculated. Thus, ASE propose that for MEDLINE document collection selecting a higher standard deviation multiplier reflects the need to account for less variability in the data; this will include the effect of small singular values and prevent ignoring important relationships.

Table 24 and Figure 14 summarize experimental results for the average standard estimator using MEDLINE document collection. It was noticed that at multiplier value of (n=1.5), k_{Est} =182 yields the best average precision, relative relevance and recall levels as can be noticed from the figures in Appendix F. ASE experimental results at (n=1.5) coincides with MEDLINE experimental results shown in Figure 12 over all possible dimensions.

Table 24: Summary of MEDLINE ASE results for various standard deviation multipliers

Standard Deviation factor in ASE (n) $ASE = \frac{\sum_{m=1}^{r-1} sv_{(m+1)} - sv_{(m)}}{r-1} + (n)S.D$	k_{Est}	***Average Precision***	***Average Relative Relevance***	***Average Recall***	***ASL***	***Average query processing time***
0	1033	0.62	1.554	0.306	1.0526	131.894
0.5	634	0.64	1.5602	0.3146	1.0543	76.144
1	338	0.6633	1.626	0.3257	1.0892	38.756
1.5	182	0.6833	1.6291	0.331	1.1268	18.632
2	103	0.6667	1.7233	0.3285	1.1331	7.254
2.5	59	0.6433	1.5947	0.3092	1.0759	3.089
3	36	0.64	1.6024	0.3044	1.0915	1.846

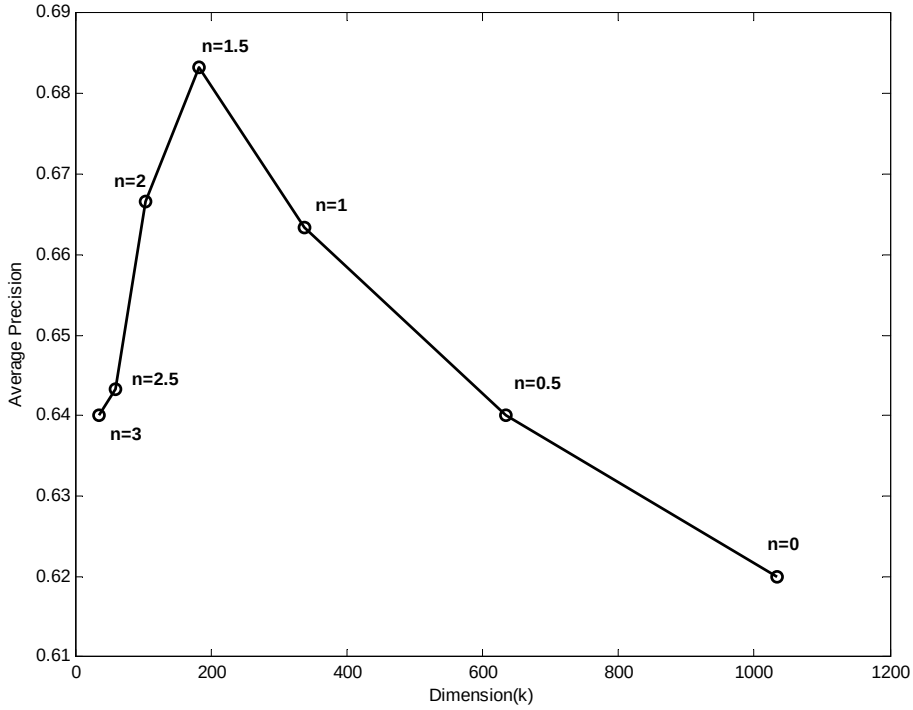


Figure 14: MEDLINE average standard estimator precision plot over a range of standard deviation multiplier's (n).

Table 25 and Figure 16 summarize experimental results for various dimensionality estimation methods with MEDLINE documents collection. For example, to account for 90% of variance then we find that k_{Est} =681, weighted model estimated k =109 with average precision of (0.660) while scree plot shown in Figure 15 estimates intrinsic dimensionality for MEDLINE at k_{Est} =203. Results over k_{Est} =203 yields average precision of (0.677) and average relative relevance of (1.116). From Table 25 and Figure 16, it is obvious that ASE performance at (n=1.5) is the best estimate for MEDLINE data intrinsic dimensionality among all other tested methods.

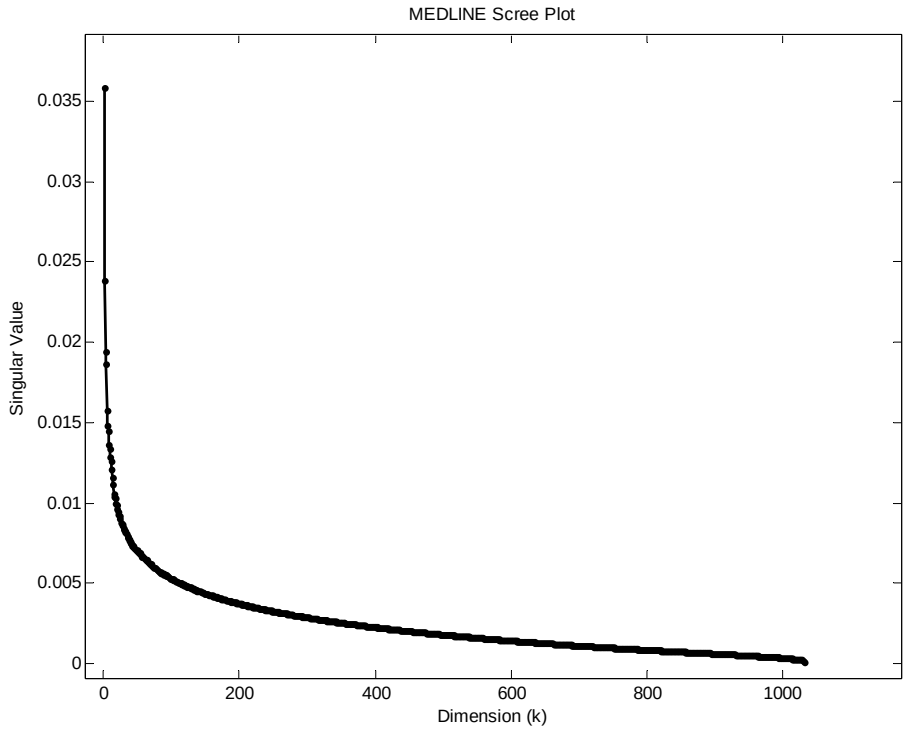


Figure 15: MEDLINE singular values scree plot

Table 25: Summary MEDLINE dimensionality estimation performance measures

Method	k_{Est}	*Average Precision*	*Average Relative Relevance*	*Average Recall*	*ASL*	*Average processing time/query*
Weighted Model	***109***	0.660	1.096	0.326	1.661	15.66
ASE (n=1.5)	***182***	0.683	1.127	0.331	1.629	34.47
Scree plot	***203***	0.677	1.116	0.333	1.662	38.85
Kaiser-Guttman	***358***	0.650	1.057	0.320	1.579	75.13
Percentage of variance (90%)	***681***	0.620	0.998	0.305	1.482	142.86

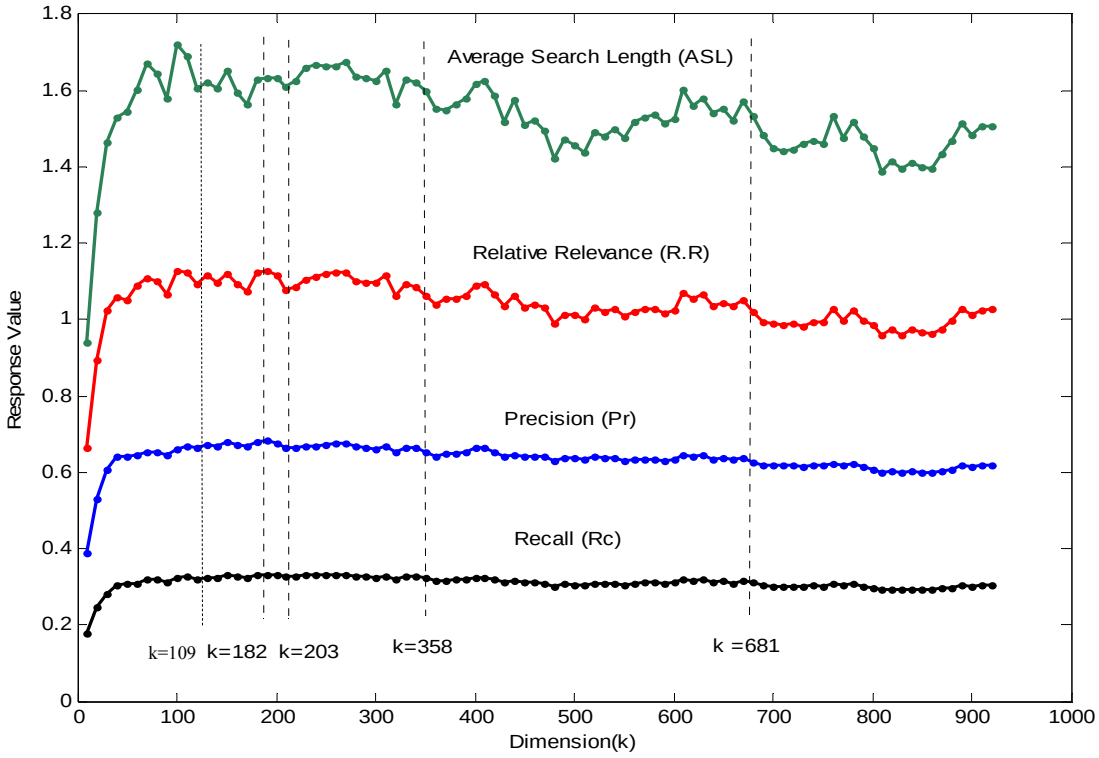


Figure 16: MEDLINE dimensionality estimation techniques performance measures

The most important result evident in all previous results is the disagreement among the various dimensionality estimation methods and performance evaluation measures. In most cases the four performance metrics were optimized at widely different dimensionalities. Overall, average search length calls for models with lower intrinsic dimensionality than do average precision, recall and relative relevance. Thus there is no clear relationship between matrix size and average search lengths estimation of matrix dimensionality. While the closer agreement between average precision, recall and relative relevance might give us the option to discount ASL because of its divergence from other metrics. Results support the need to seek a balance between different models called for by each performance evaluation criteria.

Analytical hierarchy processing performance ranking for studied dimensionality estimation techniques on MEDLINE test collection is shown in Figure 17. According to subject matter experts responses for performance measures priorities we notice that the average standard estimator results in MEDLINE outperformed all other dimensionality estimation methods followed by scree plot and the weighted model. Among the data of Table 23 the highest dimensionality reduction was found for the MEDLINE documents collection. Using the average search length measure, MEDLINE's estimated dimensionality was 90 or 8.7% of the total possible dimensions for a full rank matrix. Similar results were obtained with precision, recall and relative relevance evaluation metrics, with k_{Est} (Relative Relevance) =100 or 9.6% of full model and k_{Est} (Precision) = k_{Est} (Recall) = 150, or 14.5% of the full model.

Figure 16 shows performance measures graphs of several dimensionality estimation methods. Performance of the full-rank model is shown along with other dimensionality

estimation results, from Figures 12 and 16 it is clear that MEDLINE collection retrieval performance found with the average standard estimator technique at k_{Est} = 182 is better than all other dimensionality estimation techniques. Results suggest that LSI's reduced model improves retrieval for MEDLINE across multiple performance measures. This coincides with Deerwester et al. suggestions that MEDLINE is especially amenable to dimensionality reduction since it was constructed by a series of keyword queries. This implies that a set of well-defined concepts may be evident in the MEDLINE document collections and reflect its suitability for dimensionality reduction since results obtained by setting the model at k_{Est} (Precision) are nearly identical to those found for k_{Est} (Recall). The agreement between multiple performance metrics suggests that in the case of MEDLINE, performance evaluation metrics analysis detects the intrinsic dimensionality in the neighborhood of 200.

CRANFIELD test collection experimental results shown in Table 23 and Figure 18 indicate a disagreement between selected performance measures. We still notice the relationship of query processing time with the number of dimensions retained as indicated in Figure 19. As the number of dimensions increase, processing time increase until reaching matrix full rank where processing time is at its highest level.

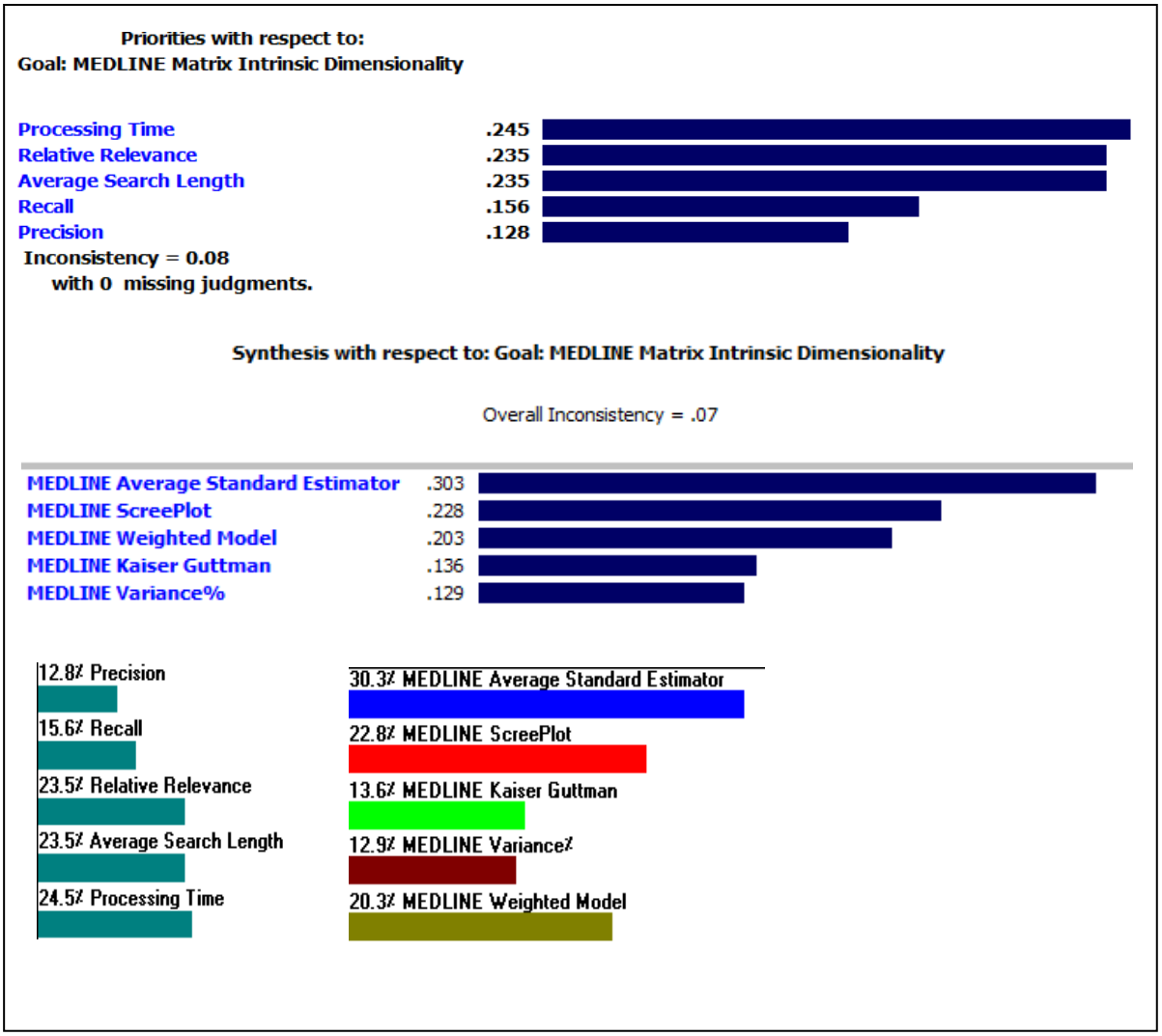


Figure 17: MEDLINE AHP performance ranking for dimensionality estimation techniques

Experimental results for CRANFIELD documents collection over all possible dimensions indicate that k_{Est}=320 will result in better performance for average precision, average recall and relative relevance as can be seen in Figure 18. Performance measures have been studied across all possible dimensions, average precision, average recall and relative relevance have close agreement at k_{Est}=320. From this point we can estimate CRANFIELD documents intrinsic dimensionality at k_{Est}=320.

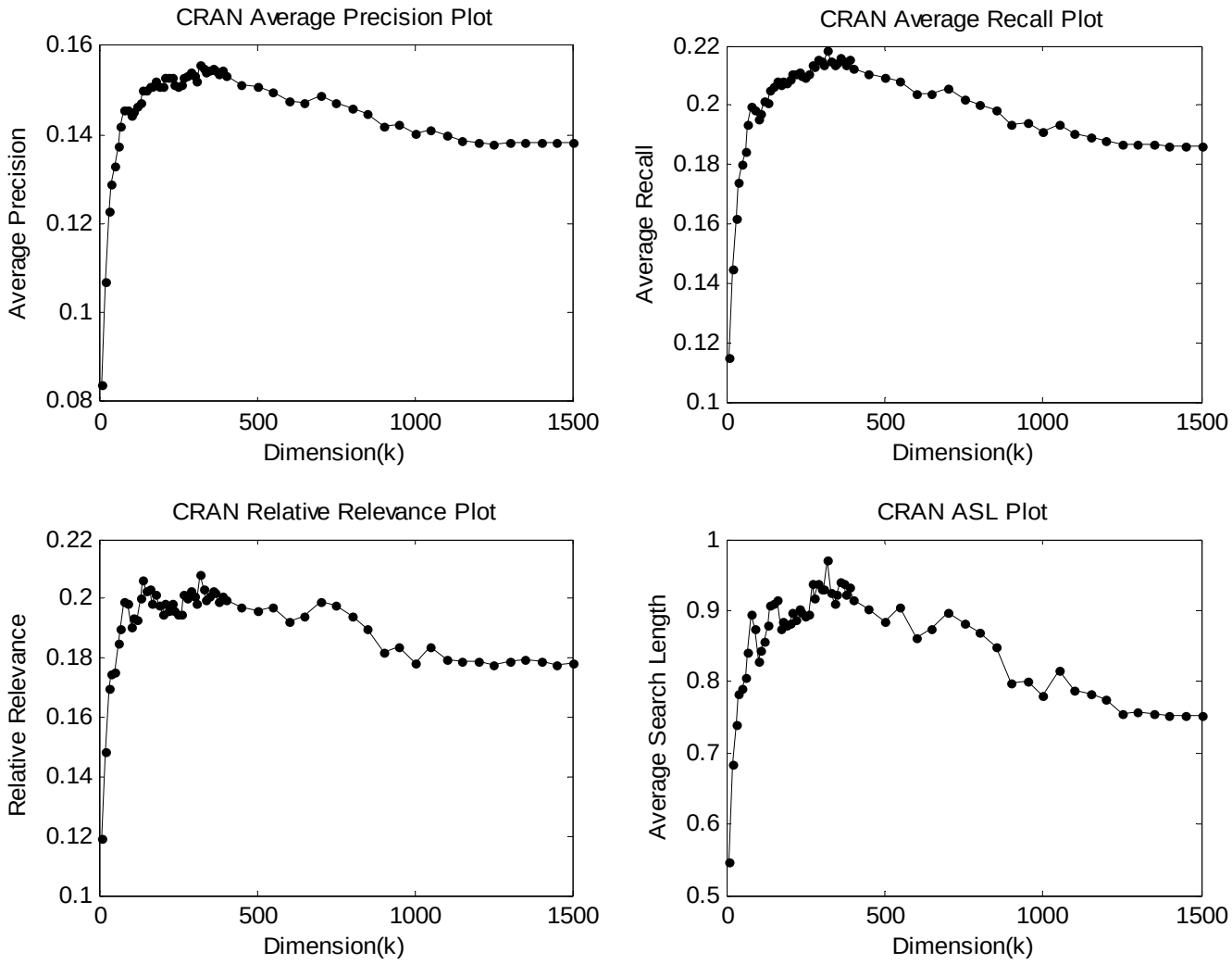


Figure 18: CRANFIELD performance results

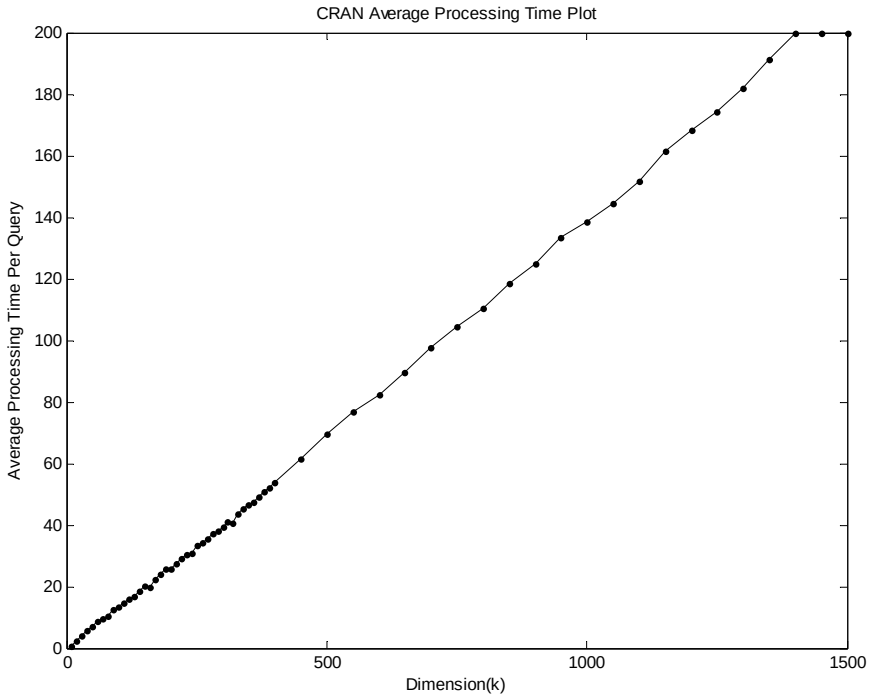


Figure 19: CRANFIELD average query processing time (Seconds)

From Table 23, using average precision and average recall performance measures in CRANFIELD, It was noticed that k_{Est} (Avg. Precision) = k_{Est} (Avg. Recall) = k_{Est} (Relative Relevance) = 320, this is 22.9% of full rank model. While average search length (ASL) performance measure performed best at k_{Est} (ASL) =100, this is 7.1% of full rank model.

ASE propose that for CRANFIELD document collection selecting a lower random noise multiplier, as shown in Table 26, reflects the need to account for more variability in the data since singular values are arranged in a descending order; this will not include the effect of smaller singular values and will ignore random relationships, since lower multiplier values will result in a decline of those factors corresponding to relatively small singular values which contain essentially random noise distracters.

Table 26 and Figure 20 summarize experimental results for the average standard estimator using CRANFIELD document collection. It was noticed that a relatively low random noise multiplier (n=1), resulted in matrix intrinsic dimensionality estimation of k_{Est}=231 with average precision of (0.1522) and average relative relevance of (0.1972). Average standard estimator at (n=1) provided the best estimation that ASE can achieve but does not completely coincides with CRANFIELD experimental results over all possible dimensions shown in Figure 18 or Appendix F which estimates k_{Est}=320.

Table 26: Summary of CRANFIELD ASE results for various standard deviation multiplier's (n)

Standard Deviation factor in ASE (n) $ASE = \frac{\sum_{m=1}^{r-1} sv_{(m+1)} - sv_{(m)}}{r-1} + (n)S.D$	k_{Est}	***Average Precision***	***Average Relative Relevance***	***Average Recall***	***ASL***	***Average query processing time (Seconds)***
0	1398	0.1384	0.1781	0.186	0.7521	251.16
0.5	515	0.15	0.1957	0.2082	0.8916	72.02
1	231	0.1522	0.1972	0.211	0.8972	28.68
1.5	110	0.1451	0.1933	0.1972	0.8438	6.722
2	56	0.1371	0.1900	0.184	0.8277	3.032
2.5	33	0.1254	0.1709	0.1666	0.7652	1.903
3	23	0.1116	0.153	0.1475	0.6771	1.414

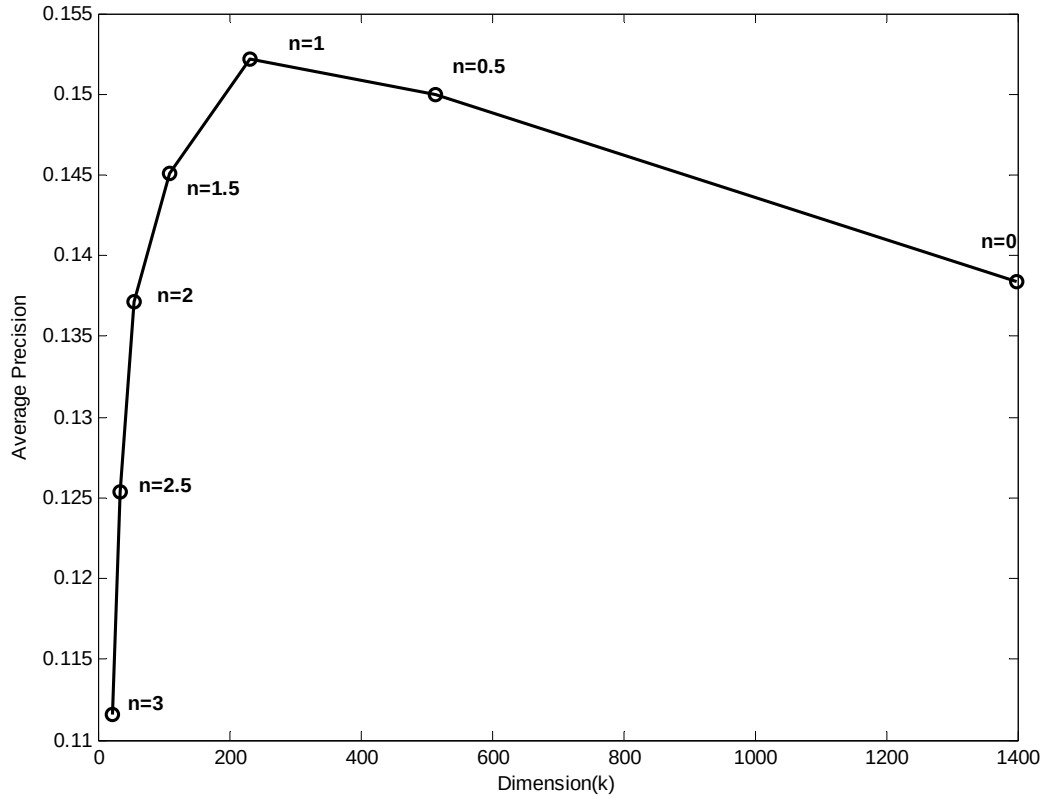


Figure 20: CRANFIELD average standard estimator precision plot over a range of standard deviation multiplier's (n).

Table 27 and Figure 22 summarize experimental results for various dimensionality estimation methods using CRANFIELD documents collection. If we want to account for 90% of the variance then we find that dimensionality was estimated at k_{Est} =804 while scree plot shown in Figure 21 estimates intrinsic dimensionality for CRANFIELD in the neighborhood of 290, results over k_{Est} =290 yields average precision of (0.154) and average relative relevance of (0.203). The weighted model estimates k at (214). From Table 27 and Figure 22, we notice that percentage of variance method when accounting for (90%) of the variance tends to overestimate data intrinsic dimensionality. We conclude that Scree plot,

weighted model and ASE at (n=1) provides better estimation of data intrinsic dimensionality for CRANFIELD documents collection than Kaiser-Guttman and Percentage of variance.

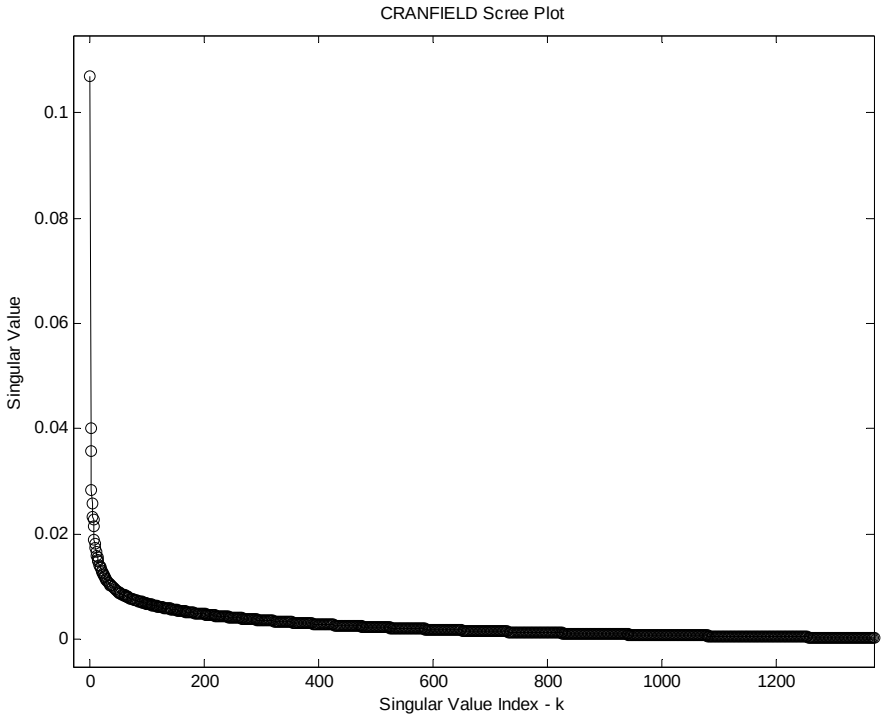


Figure 21: CRANFIELD singular values scree plot

Table 27: Summary of CRANFIELD dimensionality estimation performance measures

Method	k_{Est}	***Average Precision***	***Average Relative Relevance***	***Average Recall***	***ASL***	***Average processing time/query (Seconds)***
Weighted Model	214	0.1527	0.1984	0.2108	0.8984	27.54
ASE (n=1)	231	0.1522	0.1972	0.211	0.8972	28.68
Scree plot	290	0.154	0.203	0.215	0.938	37.881
Kaiser-Guttman	440	0.151	0.197	0.210	0.903	61.282
Percentage of variance (90%)	805	0.146	0.194	0.200	0.869	110.370

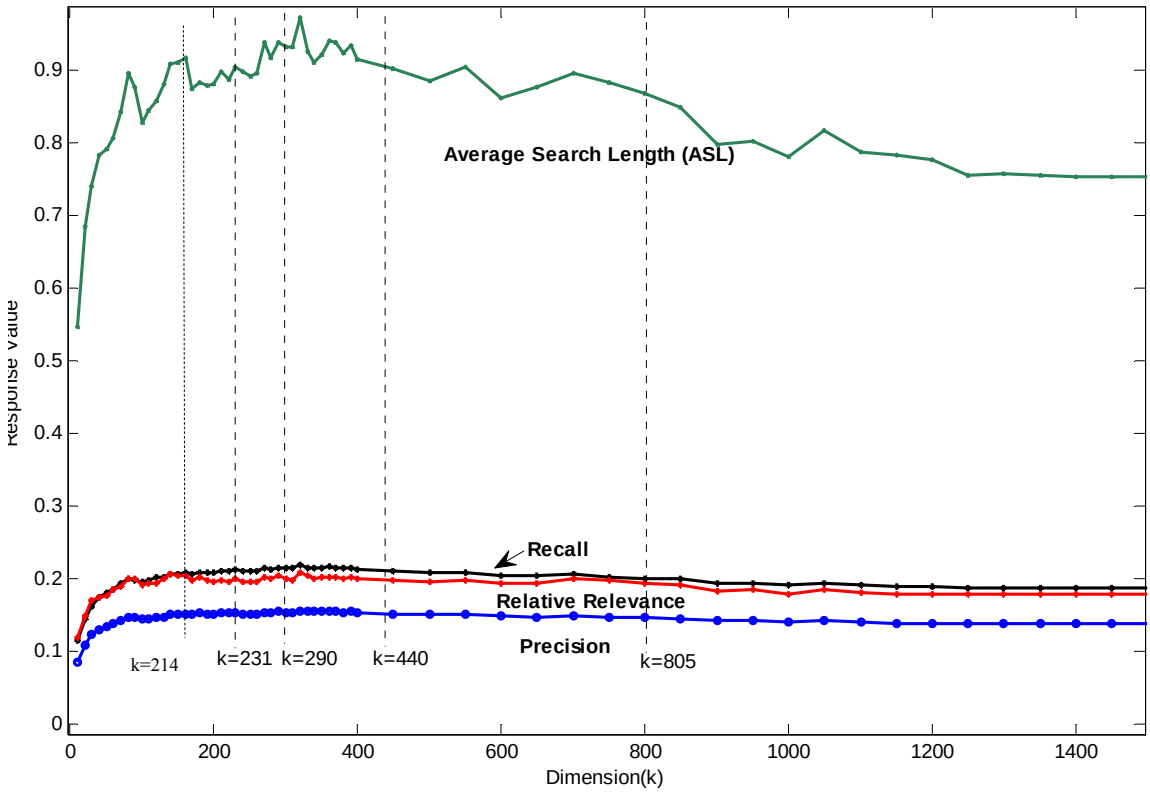


Figure 22: CRANFIELD dimensionality estimation techniques performance measures

AHP performance ranking for tested dimensionality estimation techniques on CRANFIELD test collection is shown in Figure 23. According to performance measures priorities provided by subject matter experts, it's clear that the multi-weighted model achieved the best results followed by scree plot and the average standard estimator.

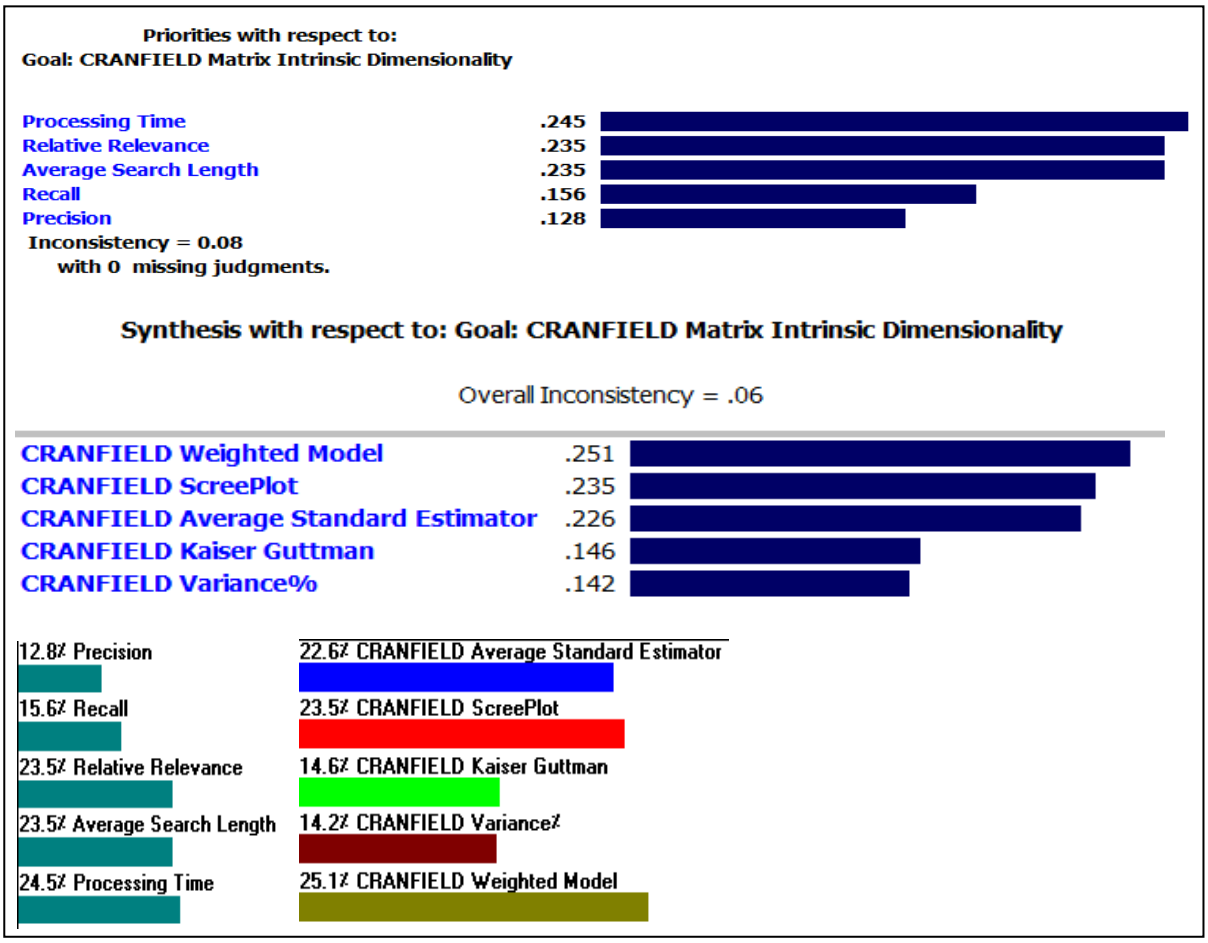


Figure 23: CRANFIELD AHP performance ranking for dimensionality estimation techniques

CISI documents collection experimental results shown in Table 23 and Figure 24 clearly indicates the strong disagreement between all performance measures. We still notice the linear relationship of query processing time with the number of dimensions retained as indicated in Figure 25. Experimental results for CISI documents collection over all possible dimensions indicate that k_{Est}=1350 will result in best average precision performance, k_{Est}=1250 will result in best average recall performance and k_{Est}=350 will result in best

average search length performance while k_{Est}=850 will result in best relative relevance performance.

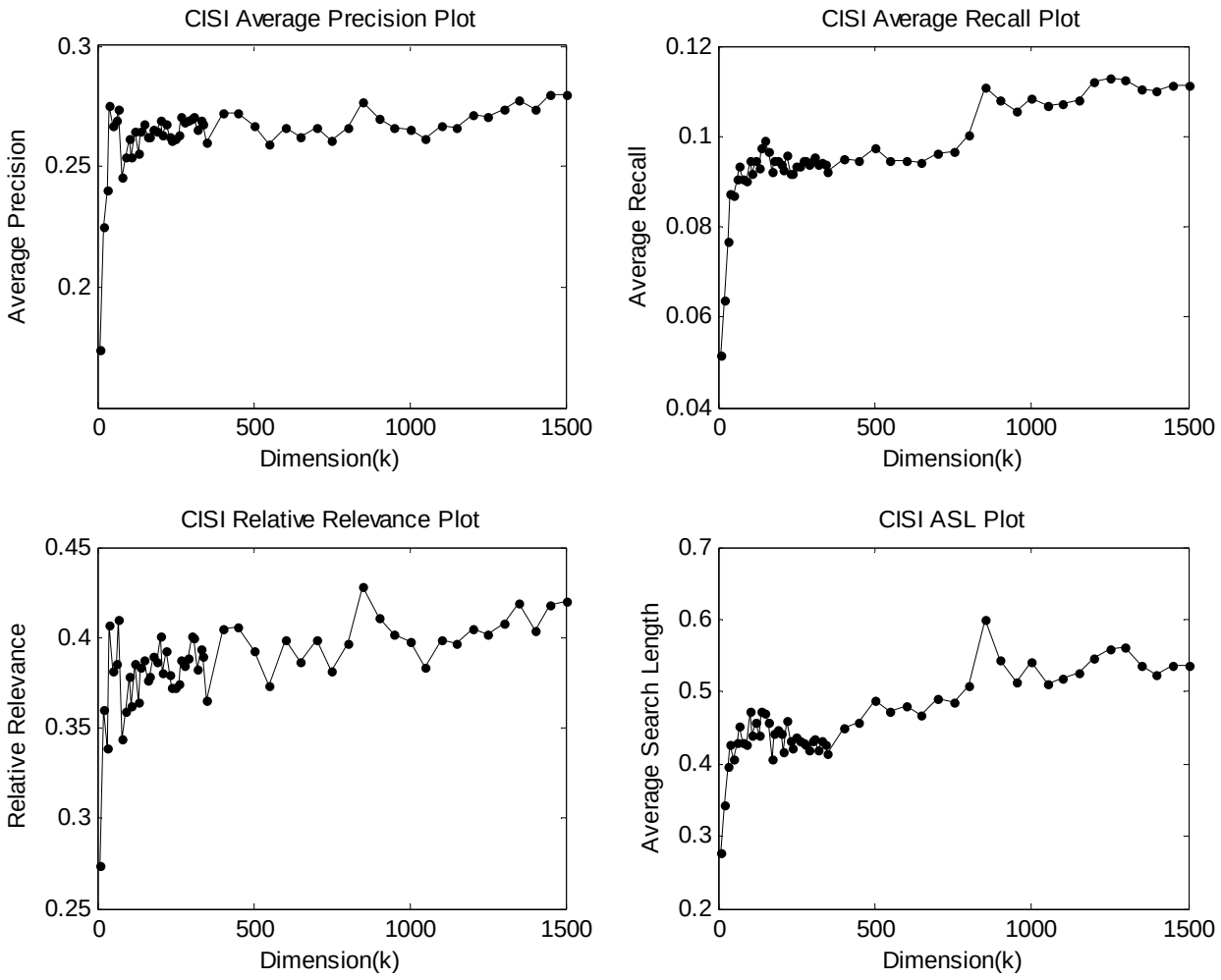


Figure 24: CISI performance results

From Table 23, considering average precision and average recall performance measures in CISI documents collection, It was noticed that there is a clear disagreement between all performance measures such that k_{Est} (Avg. Precision) ≠ k_{Est} (Avg. Recall) ≠ k_{Est} (Relative Relevance) ≠ k_{Est} (Average Search Length). Based on experimental results, it is obvious that the Average Standard Estimator (ASE) technique propose that for CISI documents

collection selecting a random noise multiplier of (0) reflects the need to account for more variability in the data since singular values are arranged in a descending order; this will include the effect of smaller singular values since lower multiplier values will result in including those factors corresponding to relatively small singular values.

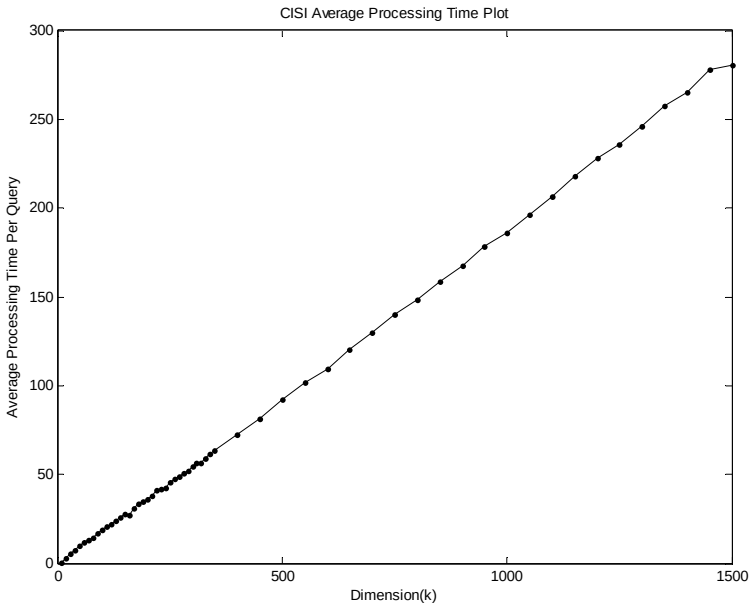


Figure 25: CISI average query processing time (Seconds)

Table 28 and Figure 26 summarize CISI experimental results for the average standard estimator. It was noticed that neglecting random noise distracters by selecting a very low multiplier results in matrix dimensionality of k_{Est} =1454 with average precision of (0.2795) and average relative relevance of (0.4184). Average standard estimator results at (n=0) is the best estimation that ASE can achieve for CISI. ASE results at (n=0) provide a good estimates of CISI documents intrinsic dimensionality and this coincides with CISI experimental results over all possible dimensions as shown in Figure 24.

Table 28: Summary of CISI ASE results for various standard deviation multiplier's (n)

Standard Deviation factor in ASE (n) $ASE = \frac{\sum_{m=1}^{r-1} sv_{(m+1)} - sv_{(m)}}{r-1} + (n)S.D$	k_{Est}	***Average Precision***	***Average Relative Relevance***	***Average Recall***	***ASL***	***Average query processing time (Seconds)***
0	1454	0.2795	0.4184	0.1114	0.537	276.54
0.5	798	0.2652	0.3970	0.0994	0.5033	139.46
1	424	0.2679	0.4051	0.0955	0.4737	69.72
1.5	229	0.2625	0.3792	0.0926	0.4374	35.17
2	121	0.2607	0.3853	0.0948	0.4556	13.21
2.5	63	0.2768	0.3857	0.0915	0.445	4.41
3	34	0.258	0.3392	0.0825	0.4224	2.36

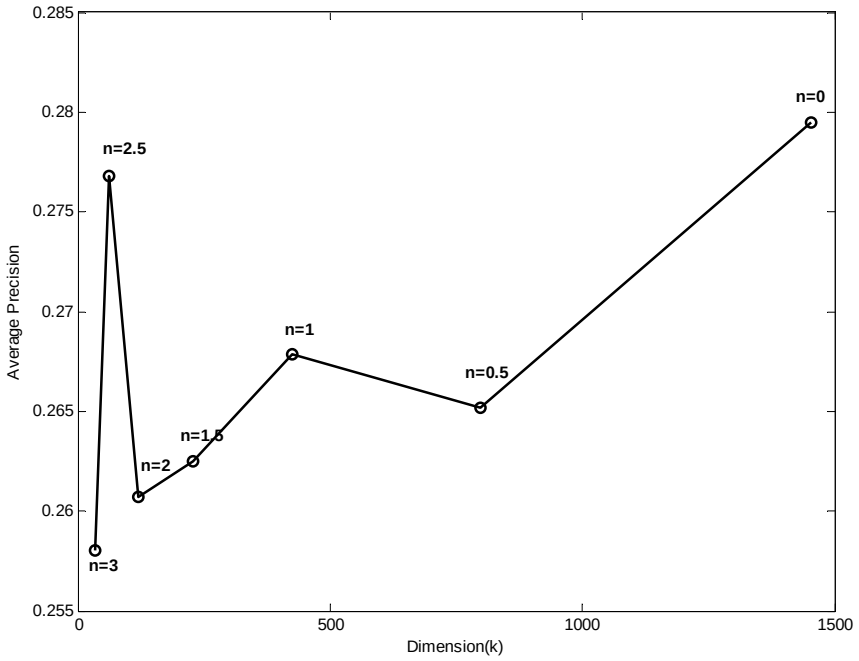


Figure 26: CISI average standard estimator precision plot over a range of standard deviation multiplier's

Table 29 and Figure 27 summarize CISI experimental results for various dimensionality estimation techniques. Experimental results for CISI document collection indicates that if we want to account for 90% of the variance then we estimate intrinsic dimensionality at 913 while scree plot shown in Figure 28 estimates intrinsic dimensionality for CISI at $k_{Est} = 600$. Performance measures calculations over k_{Est} =600 yields average precision of (0.2661) and average relative relevance of (0.3985). From Figure 27 and Appendix F for CISI test collections, we notice that all methods except ASE at (n=0) tend to underestimate CISI documents collection intrinsic dimensionality. We conclude that ASE dimensionality estimation technique at (n=0) provides better estimation of CISI intrinsic dimensionality than all other methods. Additionally, based on CISI performance over a range of random noise multipliers, ASE was able to detect irregularities at (n=2.5 and n=0) and high noise in CISI. Research results on CISI suggest that ASE would add so many benefits by eliminating noise or non relevant data in such databases.

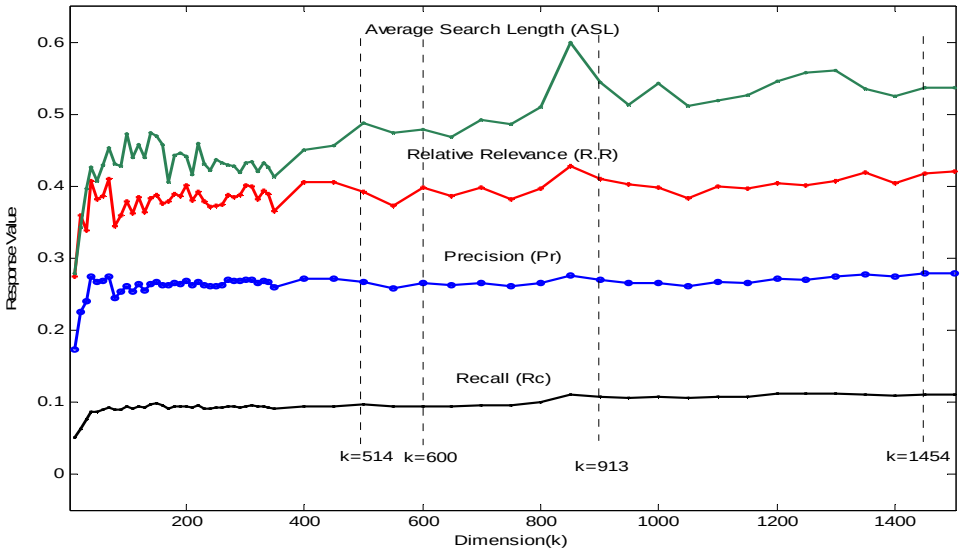


Figure 27: CISI performance response

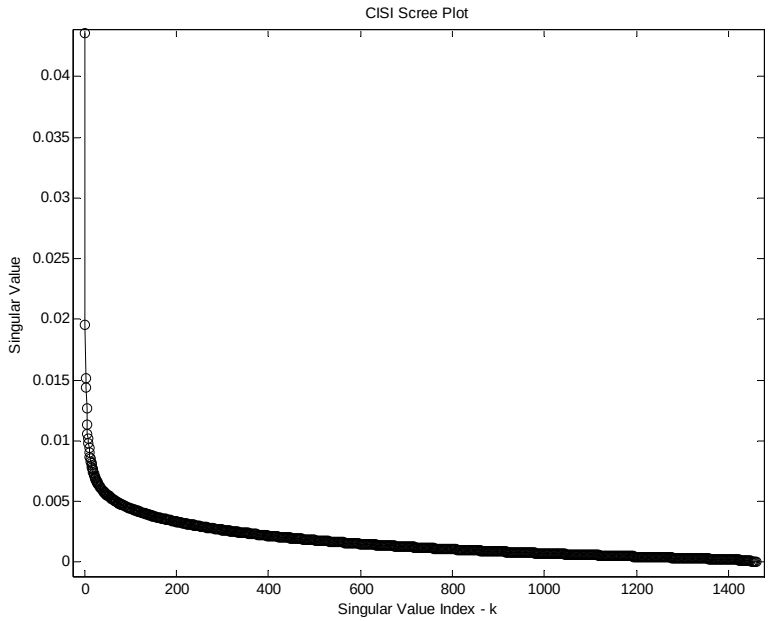


Figure 28: CISI singular values scree plot

Table 29: Summary of CISI dimensionality estimation performance measures

Method	k_{Est}	***Average Precision***	***Average Relative Relevance***	***Average Recall***	***ASL***	***Average processing time/query (Seconds)***
ASE (n=0)	***1454***	0.2795	0.537	0.1114	0.4184	276.54
Percentage of variance (90%)	***913***	0.269	0.5248	0.1062	0.4083	162.09
Kaiser-Guttman	***514***	0.267	0.3928	0.0974	0.488	92.24
Scree plot	***600***	0.2661	0.3985	0.0948	0.4792	109.59
Weighted Model	***736***	0.2616	0.3828	0.0954	0.4659	129.869

Analytical hierarchy processing performance ranking for selected dimensionality estimation techniques on CISI test collection is shown in Figure 29. According to performance measures priorities by subject matter experts, it's clear that the average standard estimator results outperformed all other dimensionality estimation techniques followed by Percentage of Variance and Kaiser Guttmann techniques.

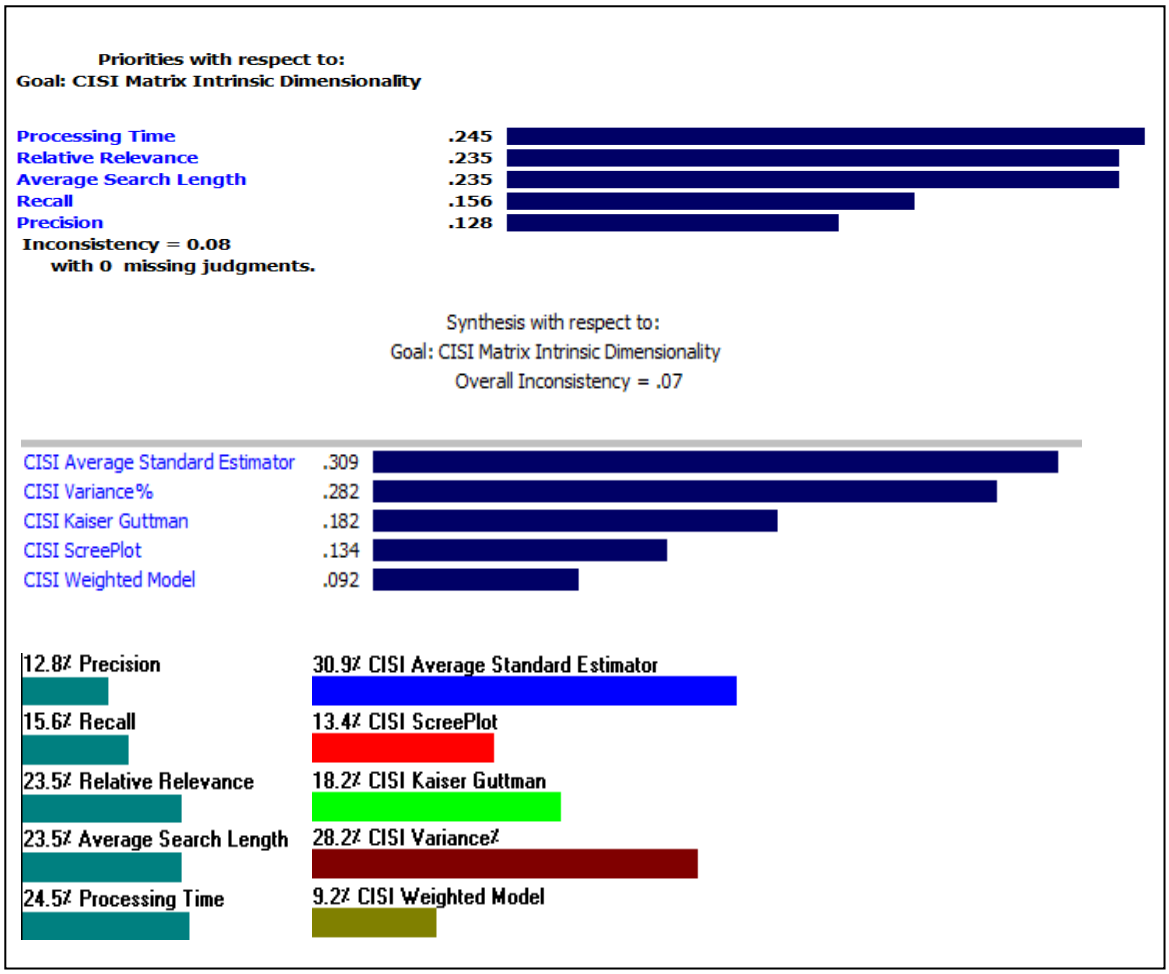


Figure 29: CISI AHP performance ranking for dimensionality estimation techniques

6.3 Summary of Results and Findings

Experimental results indicated that the average standard estimator (ASE) provided the best estimate for MEDLINE collection intrinsic dimensionality among all other dimensionality estimation techniques. Also, it was noticed that scree plot, weighted model and ASE at (n=1) provided better estimation of data intrinsic dimensionality for CRANFIELD collection than Kaiser-Guttman and Percentage of variance. For CISI documents collection, we notice that all methods except ASE tend to underestimate CISI documents collection intrinsic dimensionality. Experimental results indicates that only ASE dimensionality estimation technique at (n=0) provides a better estimation of CISI intrinsic dimensionality than all other tested methods.

As shown in Table 19, average term frequency in CRANFIELD documents collection is 57 Terms/Document, this is higher than in MEDLINE (48 Terms/Document) or CISI (46 Terms/Document). Likewise, CRANFIELD median term-document frequency is 17 Document/Term, versus MEDLINE's term-document frequency of 9 and CISI term-document frequency of 12. In other words, CRANFIELD displays more term repetition than does MEDLINE or CISI.

Latent semantic indexing advantage to CRANFIELD data may thus be attributable to the redundancy and increased term-document frequency of CRANFIELD's terms. Noticing that document collections with large number of distinct indexing terms will perform better under dimensionality reduction in LSI. Since larger termspace reflect a greater opportunity for synonymy to negatively affect retrieval results. In all experiments, smaller document collections get the greatest benefit from dimensionality reduction. Table 23 show an inclination toward increased dimensionality reduction for models with smaller

documents collections and a tendency to smaller models for document collections with greater amounts of term repetitions. Thus collections with greater repetition of terms benefited more from dimensionality reduction.

The fact that all four performance measures were not always in agreement about intrinsic dimensionality complicates the task of finding matrix intrinsic dimensionality and this reflects the need to search for better intrinsic dimensionality estimates. As discussed earlier, CISI provided little benefits by dimensionality reduction, at least so far as average precision and average recall were concerned k_{Est} (Precision) =1350 and k_{Est} (Recall) =1250. But from Table 23 it can be seen that according to the average search length (ASL), where intrinsic dimensionality was estimated at k_{Est} (ASL) =350 or 23.9% of full rank model, dimensionality reduction did improve retrieval performance for CISI for average search length performance measure only. To help understand the dynamics of dimensionality reduction, Figure 25 demonstrates reduced-rank and full-rank retrieval performance as measured by precision, recall ASL and relative relevance for the CISI data. Here dimensionality reduction provides no discernible advantage, with the precision-dimensionally reduced model and the full-rank solution showing nearly similar behavior.
On the other hand ASL dimensionally reduced model gives significantly worse results than the full-rank model. All dimensionally reduced models converge on similar performance at high levels of intrinsic dimensionality, with only the ASL dimensionally reduced model offering slight benefit.

In CISI collection it was noticed that a heavy dimensionality reduction deprives the model of important discriminatory power required to achieve good retrieval performance, In fact, the disagreement between all four performance evaluation measures for CISI and the

failure of any of them to demonstrate a strong and convincing improvement over the full-rank model by means of dimensionality reduction implies that analysis based solely on these performance measures may not be sufficient for accurately estimating intrinsic dimensionality of documents collection. This is not surprising since the average number of relevant documents per query in CISI is 50 (the highest among all other test collections), and that this collection has the largest number of total documents among other collections. Thus it might be possible that the 112 tested queries were not adequate to gather a complete and accurate estimation of CISI intrinsic dimensionality.

In Table 23 Similar but less obvious situation appears for CRANFIELD with its 225 queries, the average number of relevant documents per query is only 8, the lowest among tested document collections. It is important to mention that CRANFIELD contains a large number of documents that are not relevant to any query. These documents have been included in experimental studies. However in some studies conducted by Jiang and Littman these documents have been removed prior to analysis (Jiang et al., 2000). Non relevant documents were not removed. Since, including these documents will provide models capable of achieving better retrieval flexibility in terms of search queries, and to act as an evidence for the capabilities of performance measures to find the actual intrinsic dimensionality. Table 30 summarize my findings regarding document collections intrinsic dimensionality, these results have been discussed in detail in Section 6.2. However for each collection, Table 30 summarizes twelve measured statistics grouped into five performance measures. Rows labeled "*dimensions retained*" show the percentage of dimensions retained in the model under the specified performance evaluation metric. Rows named "*performance*

measure improved" reflect the percent of improvement over the full rank model under the specified performance measure.

Table 30: Summary of intrinsic dimensionality results

Performance evaluation measure	***MEDLINE***	***CRANFIELD***	***CISI***
Dimensions retained (Precision)	14.5%	22.9%	92.5%
Precision performance improvement	0.10	0.13	-0.01
Processing time performance improvement	124.76	159.16	20.28
Dimensions retained (Recall)	14.5%	22.9%	85.6%
Recall performance improvement	0.08	0.18	0.01
Processing time performance improvement	124.76	159.16	41.71
Dimensions retained (Relative Relevance)	9.6%	22.9%	58.2%
Relative Relevance performance improvement	0.07	0.16	0.02
Processing time performance improvement	131.83	159.16	119.21
Dimensions retained (ASL)	8.7%	7.1%	23.9%
ASL performance improvement	0.02	0.10	0.23
Processing time performance improvement	121.93	186.58	214.13

In general, average search length performance measure was associated with more dimensionality reduction than average precision, recall or relative relevance. Also, the percentage of total dimensions retained across all three test collections varies widely. Dimensionality reduction provided highest retrieval improvement for MEDLINE and CRANFIELD collections where the dimensionally reduced models improved performance greatly. The average search length performance measure indicated that CRANFIELD collection performed well at 93% dimensionality reduction, this provided 10% improvement over the full-rank model. MEDLINE average precision measure performed well at 85.5%

dimensionality reduction with 10% improvement over the full rank model. CISI data appear to respond poorly to dimensionality reduction since none of the four performance evaluation measures provided an evidence of a reduced dimensional model for CISI.

Finally, CISI appear to give no strong evidence regarding the benefits of dimensionality reduction, this might be due to the higher distribution of the number of relevant documents per queries (on average 50 relevant document/query as indicated in Table 23). CISI might have a dimensionally reduced structure which is not easy to find through tested performance measures since it was noticed that CISI had 2% improved performance over the full rank model with 41.8% dimensionality reduction with respect to relative relevance (k_{Est} (R.R) =850). Based on this, it seems that CISI database needs more queries to be able to estimate its intrinsic dimensionality.

Along all studied document collections, the five tested dimensionality estimation techniques provided clear estimates of datasets intrinsic dimensionality. Although there was a variation in those estimated values, dimensionality estimation methods were consistent in their predictions. In both MEDLINE and CRANFIELD test collections, the weighted technique followed by the average standard estimator for dimensionality estimation provided the highest dimensionality reduction with superior performance results among all other methods, while Kaiser-Guttman and percentage of variance results indicated poor overall model performance in terms selected performance measures. For CISI test collection, although the average standard estimator didn't provide the highest amount of dimensionality reduction, ASE results in terms of average precision, recall and relative relevance outperformed all other tested methods. Kaiser-Guttman and percentage of variance results were better estimates for CISI than scree plot and the weighted model.

AHP provided an excellent solution to rank all dimensionality estimation techniques according to subject matter expert's retrieval performance priorities. Table 31 summarizes dimensionality estimation techniques performance results with respect to average precision performance measure for each document collection. Table 31 indicates the direct difference of dimensions found by various estimation techniques from the dimension associated with best performance for average precision. For example the difference between intrinsic dimensionality estimation of ASE and Precision performance measure is 32 ($k_{MEDLINE}(\text{ASE}) - k_{MEDLINE}(\text{Pr}) = 32$). This indicates that ASE overestimated the precision performance measure dimension k_{Pr} for MEDLINE by 32 dimensions. Tables 32 and 33 provide similar calculations for average search length and average relative relevance where bold values indicate best performance with respect to each dimensionality estimation technique. In all tables, a value near zero indicates better estimations performance with respect to selected performance measure. Tables 31, 32 and 33 indicate that the average standard estimator and the weighted model performed well in providing good estimates for MEDLINE and CISI. This performance is not clear for CRANFIELD collection. As have been concluded before in our experiments, CRANFIELD median term-document frequency is higher than MEDLINE's and CISI term-document frequency (17 versus 9 and 12 respectively). Because of this variation in term frequency, CRANFIELD displays more term repetition than does MEDLINE or CISI.

Latent semantic indexing proved some advantages to the CRANFIELD data all of which attributable to the redundancy of CRANFIELD's terms. It was noticed previously that document collections with large number of distinct indexing terms will perform better under dimensionality reduction since larger termspace reflect a greater opportunity for synonymy

to negatively affect retrieval results, thus dimensionality reduction acts as a filter to reduce synonymy negative effects.

Table 31: Dimensionality differences for Average precision performance measure ($k_{Est} - k_{\Pr}$)

Method	***MEDLINE*** ***($k_{\Pr}$=150)***	***CRANFIELD*** ***($k_{\Pr}$=320)***	***CISI*** ***($k_{\Pr}$=1350)***
ASE	32 (n=1.5)	-89 (n=1)	104 (n=0)
Weighted Model	-41	-106	-614
Scree plot	53	-30	-750
Kaiser-Guttman	208	120	-836
Variance (90%)	531	485	-437

Table 32: Dimensionality differences for Average search length performance measure ($k_{Est} - k_{ASL}$)

Method	***MEDLINE*** ***(k_{ASL}=90)***	***CRANFIELD*** ***(k_{ASL}=100)***	***CISI*** ***(k_{ASL}= 350)***
ASE	92(n=1.5)	131(n=1)	1104(n=0)
Weighted Model	19	114	386
Scree plot	113	190	250
Kaiser-Guttman	268	340	164
Variance (90%)	591	705	563

Table 33: Dimensionality differences for average relative relevance performance measure ($k_{Est} - k_{R.R}$)

Method	***MEDLINE*** ($k_{R.R}$***=100)***	***CRANFIELD*** ($k_{R.R}$***=320)***	***CISI*** ($k_{R.R}$***= 850)***
ASE	82	-89	604
Weighted Model	9	-106	-114
Scree plot	103	-30	-250
Kaiser-Guttman	258	120	-336
Variance (90%)	581	485	63

As have been discussed earlier, the average standard estimator propose that for CISI test collection selecting a random noise multiplier of (0) reflects the need to account for the variability in the data since singular values are arranged in a descending order; this will include the effect of small singular values since lower multiplier values will result in including those factors corresponding to relatively small singular values. ASE technique was found useful in this situation since it applies a practical rationale to estimate intrinsic dimensionality.

ASE method remedy the underestimation problem of intrinsic dimensionality in all other approaches by accounting for standard deviation as an important factor to accommodate for variability in document collection characteristics and in regard to the number of documents and indexed terms. ASE assumes that those variables with deep relations have sufficient correlation and that only those relationships with high singular values are significant and should be maintained. Based on the previous discussion and experimental results over all possible dimensions, ASE improved CISI matrix intrinsic dimensionality estimation by

including the effect of both singular values magnitude of decrease and random noise distracters. Dimensionality estimation performance summary shown in Tables 34, 35 and 36 indicates that ASE provided the best estimate for MEDLINE intrinsic dimensionality among all other dimensionality estimation techniques, ASE improved precision and relative relevance by 10.2% (from 0.62 to 0.683) and 7.4% (from 1.049 to 1.127) respectively. ASE reduced MEDLINE processing time by 76% (from 145.42 to 34.47). AHP analysis indicates that ASE and the weighted model ranked among the best compared to other methods with 30.3% and 20.3% in satisfying overall model goals in MEDLINE and 22.6% and 25.1% for CRANFIELD as shown in Figure 17 and 23. The weighted model improved MEDLINE relative relevance by 4.42% (from 1.049 to 1.096), while scree plot, weighted model and ASE provided better estimation of data intrinsic dimensionality for CRANFIELD collection than Kaiser-Guttman and Percentage of variance.

Table 34: MEDLINE dimensionality estimation performance summary

			Method (MEDLINE) (n =1033), (t=145.42)				
Performance measure (% improvement from full rank model)	***K_{Est} performance***	***K_{Est} Processing time (Seconds)***	***ASE (n=1.5)***	***Weighted Model***	***Scree plot***	***Kaiser-Guttman***	***Variance (90%)***
Dimensions retained			182	109	203	358	681
Precision (K_{Est}=150)	0.62	20.88	0.683	0.660	0.677	0.650	0.620
Recall (K_{Est}=150)	0.306	20.875	0.331	0.326	0.333	0.320	0.305
Relative Relevance (K_{Est}=100)	1.0496	13.805	1.127	1.096	1.116	1.057	0.998
Average Search Length (K_{Est}=90)	1.554	23.709	1.629	1.661	1.662	1.579	1.482
Processing time			34.47	15.66	38.85	75.13	142.86

ASE dimensionality estimation technique provided a better estimation of CISI intrinsic dimensionality than all other tested methods since all methods except ASE tend to underestimate CISI documents collection intrinsic dimensionality. ASE reduced CRANFIELD processing time by 85.7% (from 199.95 to 28.68), while the Weighted model reduced CRANFIELD processing time by 86.2% (from 199.95 to 27.54). ASE improved CISI average relative relevance and average search length by 28.4% (from 0.418 to 0.537) and 22.03% (from 0.536 to 0.4184) respectively.

Table 35: CRANFIELD dimensionality estimation performance summary

			***Method (CRANFIELD) (n*=1399), (t=199.95)**				
Performance measure (% improvement from full rank model)	***K_{Est} performance***	***K_{Est} Processing time (Seconds)***	***ASE (n=1)***	***Weighted Model***	***Scree plot***	***Kaiser-Guttman***	***Variance (90%)***
Dimensions retained			231	214	290	440	805
Precision (K_{Est} =320)	0.1384	40.67	0.1522	0.1527	0.154	0.151	0.146
Recall (K_{Est} =320)	0.186	40.67	0.211	0.2108	0.215	0.210	0.200
Relative Relevance (K_{Est} =320)	0.1788	40.67	0.1972	0.1984	0.203	0.197	0.194
Average Search Length (K_{Est} =100)	0.7521	13.25	0.8972	0.8984	0.938	0.903	0.869
Processing time			28.68	27.54	37.881	61.282	110.370

Table 36: CISI dimensionality estimation performance summary

			***Method (CISI)(n*=1460), (t=277.82)**				
Performance measure (% improvement from full rank model)	***K_{Est} performance***	***K_{Est} Processing time (Seconds)***	***ASE (n=0)***	***Weighted Model***	***Scree plot***	***Kaiser-Guttman***	***Variance (90%)***
Dimensions retained			1454	736	600	514	913
Precision (K_{Est} =1350)	0.2795	257.54	0.2795	0.2616	0.2661	0.267	0.269
Recall (K_{Est} =1250)	0.111	236.11	0.1114	0.0954	0.0948	0.0974	0.1062
Relative Relevance (K_{Est} =850)	0.4184	158.61	0.537	0.3828	0.3985	0.3928	0.5248
Average Search Length (K_{Est} =350)	0.5366	63.69	0.4184	0.4659	0.4792	0.488	0.4083
Processing time			276.54	129.869	109.59	92.24	162.09

In general, analysis based on selected performance measures indicates that for each document collection there is a region of lower dimensionalities associated with improved retrieval performance. However, it was noticed that there is a clear disagreement between various performances measures on the model associated with best performance. The introduction of the weighted model and AHP analysis helped in ranking dimensionality estimation techniques and facilitates satisfying overall model goals by leveraging contradicting constrains and satisfying subject matter experts priorities. AHP analysis provided for the first time a model to help rank and compare the performance of several dimensionality estimation techniques according to overall performance. This comparison was not possible before. In all previous studies, researchers were comparing dimensionality estimation methods based on a single or multiple criteria's and neglecting all other

important metrics. Results shown in Figures 17 and Tables 34 through 39 for MEDLINE, CRANFIELD and CISI test collections indicated that the average standard estimator technique provided better results than other dimensionality estimation techniques and ranked the best among other tested methods according to AHP analysis, satisfying overall information retrieval model performance goals. AHP results and Figure 23 indicates that the weighted model not ASE provided the best estimates for CRANFIELD test collection. My explanation for this is that CRANFIELD contains a large number of documents that are not relevant to any query and acts as noise and prevented accurate dimensionality estimation for ASE. Additionally, CRANFIELD displays more term repetition than does MEDLINE or CISI. Dimensionality estimation advantage to CRANFIELD data may thus be attributable to the redundancy of CRANFIELD's terms, although this seems to be a disadvantage to the average standard estimator technique.

CHAPTER SEVEN: CONCLUSIONS

This chapter concludes experimental work and analyses covered in Chapter 6 by answering my initial research questions. Intrinsic dimensionality estimation techniques studied were very useful in providing good means for estimating documents collections intrinsic dimensionalities and to evaluate performance based on independent performance evaluation metrics. Previous research found that there is no consensus about the most effective method for estimating data intrinsic dimensionality which will provide improved overall retrieval performance. Section 7.1 in this chapter begins with coverage of my initial research questions and their theoretical significance, summarizing dimensionality estimation results and discussing each method strengths and weaknesses. Section 7.2 concludes experimental results and their implications for information retrieval. Section 7.3 describes shortcomings of this study and suggests future work on information retrieval dimensionality estimation.

7.1 Singular Value Decomposition and Dimensionality Estimation

Research covered in this study indicated that dimensionality reduction provides a better solution to information retrieval problems discussed in Chapter 1 and Chapter 2 of this research. Dimensionality reduction improved information retrieval by providing more relevant results and faster computational time, while giving reasonable accuracy in terms of precision, recall, higher relative relevance and lower average search length.

Salton's Vector Space Model (VSM) discussed in Chapter 2 treats documents as vectors in a dimensional space with inter-document similarity represented by their corresponding vector cosine (Salton et al., 1983). Documents that are about similar topics lie near each other in

Salton's vector space model. Thus information retrieval is concerned with navigating this vector space; attempting to locate regions of the vector space that contain documents relevant to specific information needs. Salton's VSM deviates from reality by assuming simplicity when VSM suggested statistical independence among terms.

Generalized Vector Space Model (GVSM) removes error from Salton's Vector Space Model (VSM) theory by including the observed term correlations as discussed in Section 1.3. Latent Semantic Indexing (LSI) removes error from the GVSM through a model based on the observed sample of the population correlation matrix. Thus LSI extends Wong GVSM by attempting to improve the model by creating a statistical model of the population correlation matrix via dimensionality reduction.

Latent Semantic Indexing (LSI) introduce the basis for a vector space by an orthogonal projection of its P-dimensional document vectors onto a k-dimensional subspace, where in LSI (k < p). Dimensionality reduction provides a systematic representation of term-document associations, similar objects are arranged by eliminating observed data over specification error (Deerwester et al., 1990). LSI is based on the singular value decomposition (SVD) of an input matrix, which was discussed in Chapter 2. Given an $n \times p$ matrix A of rank r, the singular value decomposition of A is given in Equation 7.1.1:

$$A = T \sum D' \tag{7.1.1}$$

Where T is an $n \times r$ orthogonal matrix, $\sum$ is an $r \times r$ diagonal matrix, and D is an $m \times r$ orthogonal matrix. Where matrices T and D contain the left and right singular vectors of A respectively, while the main diagonal of $\sum$ contains the singular values, which are the positive square roots of A'A and AA'. The diagonal elements of $\sum$ reflects the amount of variance of the dimensionally reduced model from the original model

Those diagonal elements of $\sum$ decrease in magnitude as i goes from 1 to rank k, this is demonstrated in Equation 7.1.2 where singular values follow a power law distribution hence the magnitude of singular values is related inversely and exponentially to the specified matrix rank k .

(7.1.2) $$\rho_1 \geq \rho_2 \geq \rho_3 \geq ..... \geq \rho_r$$

Singular values decrease in magnitude as their rank increase, because they represent the amount of variance indicated by the corresponding dimensions from the full rank model. LSI suggests that we can improve information retrieval results by neglecting those singular values with small magnitudes. Results indicated that LSI queries performance improve as the number of dimensions k increases, but this performance will decrease past a certain value of k .

Although Latent semantic indexing dimensionality reduction has proved good performance in empirical studies, the motivation behind its performance has remained largely un-formalized in previous research and literature. Several questions were un-answered such as why should a dimensionally reduced model approximation provides a better estimate of the full rank matrix!

Experimental results and analysis covered in Chapter 6 for the Average Standard Estimator (ASE) indicated that ASE was found to provide the best approximation for term-document matrix intrinsic dimensionality and better estimates than all other tested dimensionality estimation techniques. The performance of the average standard estimator supports my initial theoretical argument which states that ASE is based on the concept of terms correlation represented by singular values in SVD, thus if terms in the document collections are independent then there will be no improvement by dimensionality reduction.

ASE technique is useful since it applies a practical rationale to estimate intrinsic dimensionality.

Previous research in dimensionality reduction underestimates document collections intrinsic dimensionality. ASE method remedy the underestimation problem of intrinsic dimensionality in previous approaches by accounting for standard deviation, as an important factor to accommodate for variability in document collection characteristics and in regard to the number of documents and indexed terms. ASE assumes that variables in the document collection with deep relations have sufficient correlation and that only those relationships with high singular values are significant and should be maintained. Based on the previous discussion and experimental results, ASE improved matrix intrinsic dimensionality estimation by including the effect of both singular values magnitude of decrease and random noise distracters. Thus ASE answered one of our research questions regarding how much we need to reduce the dimensionality to derive the best estimated matrix dimensionality.

Intrinsic dimensionality or the best number of dimensions for a given corpus is thus a critical factor to the theoretical stability and success of latent semantic Indexing. Traditional matrix dimensionality estimation models do not translate easily to the unsupervised learning environment presented by information retrieval. Results in Chapter 6 confirmed that we can get better search results in terms of relevance and precision, while reducing search response time through the use of selected dimensionality reduction parameter in truncated singular value decomposition.

Since there is no consensus about the most effective method for estimating the best number of dimensions in LSI which will provide better overall retrieval performance. One of the

main objectives of this research was to develop a new and improved model to investigate the effect of various dimensionality estimation techniques on overall retireval performance. Two new techniques were introduced in this research in order to estimate matrix intrinsic dimensionality and to compare and investigate the effect of various dimensionality estimation techniques on overall search performance.

Results in Chapter 6 indicated that a system using a weighted multi-criteria performance evaluation technique resulted in better overall performance than a single criteria ranking model. Thus the weighted multi-criteria model with dimensionality reduction provides a more efficient implementation for information retrieval than what we get by using full rank model.

7.2 Findings from Experimental Data

Experimental results in Chapter 6 indicated that ASE provided the best estimate for MEDLINE collection intrinsic dimensionality among all other dimensionality estimation techniques. While scree plot, weighted model, and ASE at (n=1) provided better estimation of data intrinsic dimensionality for CRANFIELD collection than Kaiser-Guttman and percentage of variance.

Latent semantic indexing advantage to CRANFIELD data was attributable to the redundancy and increased term-document frequency of CRANFIELD terms. ASE dimensionality estimation technique at (n=0) provides a better estimation of CISI intrinsic dimensionality than all other tested methods since all methods except ASE tend to underestimate CISI document collection intrinsic dimensionality. Dimensionality reduction provided highest retrieval improvement for CRANFIELD and MEDLINE collections where

the dimensionally reduced models improved performance greatly, and where all five performance measures have been in close agreement about model intrinsic dimensionality associated with best performance. Results indicated that document collections with large numbers of distinct indexing terms will perform better under dimensionality reduction in LSI since larger termspace reflect a greater opportunity for synonymy to negatively affect retrieval results. In all experiments, smaller document collections get the greatest benefit from dimensionality reduction.

As have been discussed in Chapter 6, analysis based on selected performance measures indicate that for each document collection there is a region of lower dimensionalities associated with improved retrieval performance while there is clear disagreement between the various performance measures on the model associated with best performance. The introduction of the multi-weighted model and AHP analysis supported ranking of dimensionality estimation techniques and facilitates satisfying overall model goals by leveraging contradicting constrains and satisfying subject matter expert priorities. AHP analysis provided for the first time a model to help rank and compare performance of several dimensionality estimation techniques according to overall performance. In previous studies, researchers were comparing dimensionality estimation methods based on a single or multiple criteria and neglecting all other important metrics.

The average standard estimator technique provided better results than other dimensionality estimation techniques and ranked as the best among all other tested methods according to AHP analysis which was constructed based on expert priorities for MEDLINE and CISI.

AHP results indicate that the weighted model not ASE provided the best estimates for CRANFIELD test collection. This can be explained knowing that CRANFIELD contains a large number of documents that are not relevant to any query and acts as noise and prevents accurate dimensionality estimation for ASE. Additionally, CRANFIELD displays more term repetition than does MEDLINE or CISI. The advantage of Dimensionality estimation to CRANFIELD data may be attributable to the redundancy of CRANFIELD terms although this seems to be a disadvantage to the average standard estimator technique. Based on the Experimental results reported in this research I would suggest to revise CISI document collection by adding more queries to better estimate its dimension.

ASE served as a method to detect documents collection noise and irregular behavior through the use of the ASE plot over various noise multipliers, for example, ASE was able to detect irregular performance and high noise in CISI through the use of the random noise multiplier at n=0 and n=2.5. Thus, based on experimental results, this research suggests the use of ASE as a tool to be used in the detection of noise and non-relevant documents in such databases. Also, results clearly mark the importance of considering the random noise multiplier as a performance measure to study and evaluate information retrieval systems performance in estimating intrinsic dimensionality.

This research provided the evidence, which supports that: a system using a weighted multi-criteria performance evaluation technique resulted in better overall performance than a single criteria ranking model. Thus the weighted multi-criteria model with dimensionality reduction provides a more efficient implementation for information retrieval than what we get by using full rank model.

7.3 Study Limitations and Future Work

This section covers research limitation and provides suggestions for future work. Additionally, a couple of issues left open for future research will be discussed. One of the important issues that needs to be addressed in future research is the number of test collections and the their associated characteristics such as size, number of terms per document and all other matrix characteristics mentioned previously in this research. This research tested three document collections with distinct characteristics and qualities. However, future research should study larger document collections form either supervised or unsupervised learning environments such as large data libraries and compare results that we got for each one of them.

The numbers of performance evaluation measures were restricted to average precision, recall, relative relevance, average search length and time. Future research should study performance evaluations measures and introduce other possible candidates, such as the random noise multiplier introduced in this research, which can better estimate matrix intrinsic dimensionality and improve over the multi-weighted model results. The random noise multiplier acts as a method to detect irregularities in documents collections and possibly as a method to detect non relevant documents which acts as noise. Another important aspect is the number of tested dimensionality estimators and the number of dimensionality reduction techniques, this research studied five techniques for dimensionality estimation based on the singular value decomposition due to wide acceptance in the IR community and good performance. It would be interesting to see what will be the results under other dimensionality estimation techniques or other dimensionality reduction techniques such as the Independent component analysis and multi-dimensional scaling.

APPENDIX A: SMART STOP LIST

A'S ABLE ABOUT ABOVE ACCORDING ACCORDINGLY ACROSS ACTUALLY AFTER AFTERWARDS AGAIN AGAINST AIN'T ALL ALLOW ALLOWS ALMOST ALONE ALONG ALREADY ALSO ALTHOUGH ALWAYS AM AMONG AMONGST AN AND AN- OTHER ANY ANYBODY ANYHOW ANYONE ANYTHING ANYWAY ANYWAYS ANYWHERE APART APPEAR APPRECIATE APPROPRIATE ARE AREN'T AROUND AS ASIDE ASK ASKING ASSOCIATED AT AVAILABLE AWAY AWFULLY B BE BE- CAME BECAUSE BECOME BECOMES BECOMING BEEN BEFORE BEFOREHAND BEHIND BEING BELIEVE BELOW BESIDE BESIDES BEST BETTER BETWEEN BEYOND BOTH BRIEF BUT BY C C'MON C'S CAME CAN CAN'T CANNOT CANT CAUSE CAUSES CERTAIN CERTAINLY CHANGES CLEARLY CO COM COME COMES CONCERNING CONSEQUENTLY CONSIDER CONSIDERING CONTAIN CONTAINING CONTAINS CORRESPONDING COULD COULDN'T COURSE CURRENTLY D DEFINITELY DESCRIBED DESPITE DID DIDN'T DIFFERENT DO DOES DOESN'T DOING DON'T DONE DOWN DOWNWARDS DURING E EACH EDU EG EIGHT EITHER ELSE ELSEWHERE ENOUGH ENTIRELY ESPECIALLY ET ETC EVEN EVER EVERY EVERYBODY EVERYONE EVERYTHING EVERYWHERE EX EXACTLY EXAMPLE EXCEPT F FAR FEW FIFTH FIRST FIVE FOLLOWED FOLLOWING FOLLOWS FOR FORMER FORMERLY FORTH FOUR FROM FURTHER FURTHERMORE G GET GETS GETTING GIVEN GIVES GO GOES GOING GONE GOT GOTTEN GREETINGS H HAD HADN'T HAPPENS HARDLY HAS HASN'T HAVE HAVEN'T HAVING HE HE'S HELLO HELP HENCE HER HERE HERE'S HEREAFTER HEREBY HEREIN HEREUPON HERS HERSELF HI HIM HIMSELF HIS HITHER HOPEFULLY HOW HOWBEIT HOWEVER I I'D I'LL I'M I'VE IE IF IGNORED IMMEDIATE IN INASMUCH INC INDEED INDICATE INDICATED INDICATES INNER INSOFAR INSTEAD INTO INWARD IS ISN'T IT IT'D IT'LL IT'S ITS ITSELF J JUST K KEEP KEEPS KEPT KNOW KNOWS KNOWN L LAST LATELY LATER LATTER LATTERLY LEAST LESS LEST LET LET'S LIKE LIKED LIKELY LITTLE LOOK LOOKING LOOKS LTD M MAINLY MANY MAY MAYBE ME MEAN MEANWHILE MERELY MIGHT MORE MOREOVER MOST MOSTLY MUCH MUST MY MYSELF N NAME NAMELY ND NEAR NEARLY NECESSARY NEED NEEDS NEITHER NEVER NEVERTHELESS NEW NEXT NINE NO NOBODY NON NONE NOONE NOR NORMALLY NOT NOTHING NOVEL NOW NOWHERE O OBVIOUSLY OF OFF OFTEN OH OK OKAY OLD ON ONCE ONE ONES ONLY ONTO OR OTHER OTHERS OTHERWISE OUGHT OUR OURS OURSELVES OUT OUTSIDE OVER OVERALL OWN P PARTICULAR PARTICULARLY PER PERHAPS PLACED PLEASE PLUS POSSIBLE PRESUMABLY PROBABLY PROVIDES Q QUE QUITE QV R RATHER RD RE REALLY REASONABLY REGARDING REGARDLESS REGARDS RELATIVELY RESPECTIVELY RIGHT S SAID SAME SAW SAY SAYING SAYS SECOND SECONDLY SEE SEEING SEEM SEEMED SEEMING SEEMS SEEN SELF SELVES SENSIBLE SENT SERIOUS SERIOUSLY SEVEN SEVERAL SHALL SHE SHOULD SHOULDN'T SINCE SIX SO SOME SOMEBODY SOMEHOW SOMEONE SOMETHING SOMETIME SOMETIMES SOMEWHAT SOMEWHERE SOON SORRY SPECIFIED SPECIFY SPECIFYING STILL SUB SUCH SUP SURE T T'S TAKE TAKEN TELL TENDS TH THAN THANK THANKS THANX THAT THAT'S THATS THE THEIR THEIRS THEM

THEMSELVES THEN THENCE THERE THERE'S THEREAFTER THEREBY THEREFORE THEREIN THERES THEREUPON THESE THEY THEY'D THEY'LL THEY'RE THEY'VE THINK THIRD THIS THOROUGH THOROUGHLY THOSE THOUGH THREE THROUGH THROUGHOUT THRU THUS TO TOGETHER TOO TOOK TOWARD TOWARDS TRIED TRIES TRULY TRY TRYING TWICE TWO U UN UNDER UNFORTUNATELY UNLESS UNLIKELY UNTIL UNTO UP UPON US USE USED USEFUL USES USING USUALLY UUCP V VALUE VARIOUS VERY VIA VIZ VS W WANT WANTS WAS WASN'T WAY WE WE'D WE'LL WE'RE WE'VE WELCOME WELL WENT WERE WEREN'T WHAT WHAT'S WHATEVER WHEN WHENCE WHENEVER WHERE WHERE'S WHEREAFTER WHEREAS WHEREBY WHEREIN WHEREUPON WHEREVER WHETHER WHICH WHILE WHITHER WHO WHO'S WHOEVER WHOLE WHOM WHOSE WHY WILL WILLING WISH WITH WITHIN WITHOUT WON'T WONDER WOULD WOULDN'T X Y YES YET YOU YOU'D YOU'LL YOU'RE YOU'VE YOUR YOURS YOURSELF YOURSELVES Z ZERO ZUELZER

APPENDIX B: ASE EXAMPLE RESULTS

Table: Kaiser-Guttman dimensionality estimation results for ASE example

Kaiser-Guttman Analysis (k=358)					
Query	***Precision***	***Recall***	***ASL***	***Relative Relevance***	***Processing time***
1	0.9	0.2432	1.3514	0.9625	3.8064
2	0.5	0.3125	1.125	0.7498	3.5568
3	0.7	0.3182	1.4091	0.8031	3.6504
4	0.6	0.2609	1.0435	0.6861	3.744
5	0.9	0.3462	1.9231	1.2728	4.1964
6	0.7	0.5385	2.6154	2.1592	3.6504
7	0.6	0.4	1.6667	1.0008	3.6348
8	0.4	0.3636	1.7273	0.6187	3.7752
9	0.4	0.3636	1.7273	0.4123	3.7908
10	0.6	0.25	1.3333	0.2481	3.9312
11	0.7	0.3889	2.3333	0.5311	3.8532
12	0.7	0.7778	4.2222	1.1347	3.6348
13	1	0.4762	2.619	1.3262	3.7752
14	0.7	0.4375	2.125	0.8291	3.6504
15	0.8	0.2759	1.3448	0.8171	3.6972

Table: ASE dimensionality estimation results for ASE example

Average Standard Estimator Analysis (k=182)					
Query	***Precision***	***Recall***	***ASL***	***Relative Relevance***	***Processing time (Seconds)***
1	**0.9**	**0.2432**	**1.3784**	**1.0033**	**1.638**
2	**0.6**	**0.375**	**1.5625**	**0.9208**	**1.6224**
3	**0.7**	**0.3182**	**1.3636**	**0.9877**	**1.6224**
4	**0.8**	**0.3478**	**1.8696**	**0.7893**	**1.7784**
5	**1**	**0.3846**	**2.1154**	**1.6299**	**1.8096**
6	**0.8**	**0.6154**	**3.0769**	**2.6535**	**1.6536**
7	**0.6**	**0.4**	**1.4667**	**1.2266**	**1.6692**
8	**0.4**	**0.3636**	**1.0909**	**0.7378**	**1.5444**
9	**0.4**	**0.3636**	**1.0909**	**0.5193**	**2.1684**
10	**0.6**	**0.25**	**1.4167**	**0.2101**	**1.6848**
11	**0.7**	**0.3889**	**2**	**0.6247**	**1.794**
12	**0.6**	**0.6667**	**3**	**1.0717**	**1.6848**
13	**1**	**0.4762**	**2.619**	**1.5645**	**1.7784**
14	**0.8**	**0.5**	**2.8125**	**1.0146**	**1.6848**
15	**0.8**	**0.2759**	**1.3103**	**0.8595**	**1.716**

Table: Scree Plot dimensionality estimation results for ASE example

Scree Plot Analysis (k=203)					
Query	***Precision***	***Recall***	***ASL***	***Relative Relevance***	***Processing time (Seconds)***
1	0.9	0.2432	1.3514	1.0039	1.8876
2	0.6	0.375	1.5625	0.8908	1.9188
3	0.7	0.3182	1.4545	0.9928	2.0592
4	0.7	0.3043	1.3913	0.7633	1.9032
5	0.9	0.3462	1.7692	1.5168	1.95
6	0.8	0.6154	3.1538	2.5867	1.95
7	0.6	0.4	1.4	1.2202	2.106
8	0.4	0.3636	1.1818	0.7189	1.8252
9	0.4	0.3636	1.1818	0.5085	1.716
10	0.5	0.2083	1.0833	0.2041	1.9968
11	0.7	0.3889	1.8889	0.6146	1.9188
12	0.7	0.7778	4	1.207	1.9344
13	1	0.4762	2.619	1.5467	2.028
14	0.7	0.4375	2.1875	0.9524	1.9968
15	0.8	0.2759	1.3448	0.8558	1.95

APPENDIX C: MATLAB CODE FOR ASE EXAMPLE

```
% This Matlab Code was written by: Tareq Ahram (PhD research).
% Date: August 5, 2008
%--------------------------------------------------------------------
% MEDLINE Queries Document Collection Performance Measure for 15 queries.
% performance measures based on the first 10 most relevant documents
% returned

% Load MEDLINE data
load('MED_Original.mat');
   disp(['Some statistics about the MED data collection:'])
   disp(['  Number of rows in term-by-document matrix A is
',int2str(m),'.' ])
   disp(['  Number of columns in term-by-document matrix A is
',int2str(n),'.' ])
   disp(['  Number of nonzeros in term-by-document matrix A is
',int2str(nnz),'.' ])

 % ASE Estimation
data=A;
[b,c] = size(data);
% data normalization.
datac = data - repmat(sum(data)/b,b,1);
%Find the covariance matrix.
covm = cov(datac);
[eigvec,eigval] = eigs(covm,c);
% find singular values for the first 1032 (k<n) row and column
eigval = diag(eigval);  % extract the diagonal elements
% Calculation of Distances
for g=1:c-1
val(g)=abs(eigval(g+1)-eigval(g));
end
%calculate singular values standard deviation
stdev=std(eigval);
% calculate singular values average distance
k=Kest,
    [U,S,V]=svds(A,k);
    'ALERT!: MEDLINE Matrix Dimension Change'
    'Loading MEDLINE Data for Selected K Completed Successfully'
% Construct Empty matrix to save each loop results
 MEDresult=zeros(15,5);
 numreturn=10;

for qnumber=1:15
    if (qnumber == 1)
      qterms=[1197  2481    2482    2648    3007    3008    5706];
    elseif (qnumber==2)
      qterms=[609   610 770 1007    1008    2096    2800    2801    3294
3295    3705    3762    4106    4107    4473 ];
    elseif (qnumber==3)
      qterms=[645   646 647 1693    3114    3115    3311    3314];
    elseif (qnumber==4)
      qterms=[5404  5405    1208    1210    3114    3115    645 647 646
3484    3483];
    elseif (qnumber==5)
      qterms=[1190  1998    1993    79  82  3925    3924    3923    3926
529 530 3541    3029    3028    2030    2031    2029 ];
```

```
     elseif (qnumber==6)
       qterms=[5694  4797    1293    1294    3604    426 338 337 4462];
     elseif (qnumber==7)
       qterms=[4323  2357    4718    4719    3154    3153    1440    1437
1439    3837    1672    5104    5105    5529    1761    282 283 284 3840
5355    5354    5293    4644    4314    4315 ];
     elseif (qnumber==8)
       qterms=[1660  1665    1607    1608    615 616 617 3202    294 295
4974    4975    1660    4873    4874    786 ];
     elseif (qnumber==9)
       qterms=[2689  2690    2545    2546    2357    5208    5209    3501
2350    2757    2707    2705    1523    1524 ];
     elseif (qnumber==10)
       qterms=[3483  3484    2604    2588    2603    2602  ];
     elseif (qnumber==11)
       qterms=[609   5633    5632    5054    2481    2482    637 4208
4209    3483    3484 ];
     elseif (qnumber==12)
       qterms=[1660  1665    506 5262    3117    1811    4443    4500
3013    3012 ];
     elseif (qnumber==13)
       qterms=[514   5139    3872    2204    2203    4433    5454    5453];
     elseif (qnumber==14)
       qterms=[4500  252 987 5522    1665    1660    5054    5053    1025
5324    5329    2928    2929    1523    1524    3493    3492    5246
4769    4768    4066    4065    2800  ];
     elseif (qnumber==15)
       qterms=[2457  2384    5735    5736    341 342 3234    421 2219
5246    185 186 2800 ];
     end
   q=zeros(m,1);
   for i=1:size(qterms,2)
     q(qterms(i))=1;
   end
% query processing time
   tic=cputime;
% Document relevance calculation
normq=norm(q,2);
for j=1:n
   rowiofV=V(j,:);
   s=S*(rowiofV)';
   angle(j)=(s'*(U'*q))/(norm(s,2)*normq);
end
calcanglestime=cputime-tic;
[sortedangle,index]=sort(angle);

% MEDLINE data
     if qnumber==1
       % MED query 1
       reldocs=[13  14  15  72  79  138 142 164 165 166 167 168 169 170
171 ...
            172  180 181 182 183 184 185 186 211 212 499 500 501 502 503
504 ...
            506  507 508 510 511 513];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
       precision=sum(kset)/numreturn;
     elseif qnumber==2
```

```
      % MED query 2
      reldocs=[80  90  162 187 236 237 258 289 290 292 293 294 296 300
301 ...
          303];
      kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==3
      % MED query 3
      reldocs=[59  62  67  69  70  71  73  78  81  160 163 230 231 232
233 ...
          234  276 277 279 282 283 287];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==4
      % MED query 4
      reldocs=[93  94  96  141 173 174 175 176 177 178 207 208 209 210
259 ...
          396  397 399 400 404 405 406 408];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==5
      % MED query 5
      reldocs=[1   2   4   5   6   7   8   9   10  11  12  158 159 188
304 ...
          305  306 307 325 326 327 329 330 331 332 333];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==6
      % MED query 6
      reldocs=[112 115 116 118 122 238 239 242 260 309 320 321 323];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==7
      % MED query 7
      reldocs=[92  121 189 247 261 382 385 386 387 388 389 390 391 392
393];
      kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==8
      % MED query 8
      reldocs=[52  60  61  123 190 251 262 263 264 265 266];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==8
      % MED query 9
      reldocs=[30  31  53  56  57  64  83  84  89  124 125 126 192 252
253 ...
          267  268 269 270 271 272 273 409 412 415 420 421 422];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
    elseif qnumber==10
      % MED query 10
      reldocs=[54  55  58  152 153 154 155 254 255 256 257 529 531 532
533 ...
          534  535 537 538 539 540 541 542 543];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
      precision=sum(kset)/numreturn;
```

```
     elseif qnumber==11
       % MED query 11
       reldocs=[32  63  66  148 150 225 226 228 229 440 441 444 445 446
447 ...
           448  451 452];
        kset=ismember(index(n:-1:n-numreturn+1),reldocs);
       precision=sum(kset)/numreturn;
     elseif qnumber==12
       % MED query 12
       reldocs=[16  17  19  20  193 364 365 366 367];
        kset=ismember(index(n:-1:n-numreturn+1),reldocs);
       precision=sum(kset)/numreturn;
     elseif qnumber==13
       % MED query 13
       reldocs=[21  22  143 144 145 146 194 195 196 197 198 199 470 471
474 ...
           475  477 478 479 481 483];
        kset=ismember(index(n:-1:n-numreturn+1),reldocs);
       precision=sum(kset)/numreturn;
     elseif qnumber==14
       % MED query 14
       reldocs=[23  24  25  26  28  29  454 455 456 457 459 461 463 466
467 ...
           468];
        kset=ismember(index(n:-1:n-numreturn+1),reldocs);
       precision=sum(kset)/numreturn;
    elseif qnumber==15
       % MED query 15
       reldocs=[33  34  101 102 104 105 107 109 110 140 215 216 218 219
220 ...
           222  349 350 351 352 353 355 356 357 358 359 361 362 363];
       kset=ismember(index(n:-1:n-numreturn+1),reldocs);
       precision=sum(kset)/numreturn;
     end
     % performance measure calculation
     precision=(sum(kset)/numreturn);
     recall=(sum(kset)/(size(reldocs,2)));
     asl=sum(find(kset==1))/size(reldocs,2);
     relrelevance= ((sum(kset.*sortedangle(n:-1:n
numreturn+1)))/sqrt(sum(kset))*sqrt(sum(sortedangle(n:-1:n-
numreturn+1))));
     calcanglestime=calcanglestime;

   % Display Query Number
     Query=qnumber,
   % Display Query Performance
     QueryPerf=[precision,recall,asl,relrelevance,calcanglestime],
      MEDresult(qnumber,1) = precision;
      MEDresult(qnumber,2) = recall;
      MEDresult(qnumber,3) = asl;
      MEDresult(qnumber,4) = relrelevance;
      MEDresult(qnumber,5) = calcanglestime;
end
    '1)Precision 2)Recall 3)ASL 4)Relative Relevance  5)Time'
     MEDresult,
    % Show Average Performance measure result for selected K value
     Averageprecision=(mean(MEDresult(1:15,1))),
```

```
Averagerecall=(mean(MEDresult(1:15,2))),
Averageasl=(mean(MEDresult(1:15,3))),
Averagerelrelevance=(mean(MEDresult(1:15,4))),
Averagetime=(mean(MEDresult(1:15,5))),
```

APPENDIX D: INFORMED CONSENT AND QUESTIONNAIRE

Informed Consent

Researchers at the University of Central Florida (UCF) study many topics. To do this we need the help of people who agree to take part in a research study. You are being invited to take part in a research study which will include about four subject matter experts. You can ask questions about the research. You can read this form and agree to take part right now, or take the form home with you to study before you decide. You will be told if any new information is learned which may affect your willingness to continue taking part in this study. You have been asked to take part in this research study because you are a researcher in the field of Optimization or Information Retrieval. You must be 18 years of age or older to be included in the research study and sign this form.

The person doing this research is Tareq Ahram, a PhD candidate in the Industrial Engineering Department at the University of Central Florida. Because the researcher is a PhD student he is being guided by Dr. Pamela McCauley-Bush, a UCF faculty supervisor in the department of Industrial Engineering.

Study title: The Multi-criteria Decision Weighted model to enhance information retrieval and search engines performance.

Purpose of the research study: The purpose of this study is to participate as a Subject Matter Expert (SME) to decide on information retrieval priorities for The Multi-criteria Decision Weighted model designed to enhance information retrieval and search engines performance.

What you will be asked to do in the study: After reading the consent form, you will be presented with short questions to complete electronically. As you work through the list of questions you will select the answers that best represent your preference and priorities.

Voluntary participation: You have been selected to participate in this study as one of four participants with expertise in the Information retrieval and decision analysis research. You should take part in this study only because you want to. There is no penalty for not taking part, and you will not lose any benefits. You have the right to stop at any time. Just tell the researcher or a member of the research team that you want to stop. You will be told if any new information is learned which may affect your willingness to continue taking part in this study.

Location: Study will be conducted by e-mail

Time required: This study will take approximately (10) minutes to complete.

Risks: There are no expected risks for taking part in this study. You do not have to answer every question or complete every task. You will not lose any benefits if you skip questions or tasks. You do not have to answer any questions that make you feel uncomfortable.

Benefits:

As a research participant you will not benefit directly from this research, besides learning more about how research is conducted.

Compensation or payment:

There is no compensation or other payment to you for taking part in this study.

Confidentiality: Your identity will be kept confidential. The researcher will make every effort to prevent anyone who is not on the research team from knowing that you gave us information, or what that information is. For example, your name will be kept separate from the information you give, and these two things will be stored in different places.

Your information will be assigned a code number. The list connecting your name to this number will be kept in a locked file cabinet. When the study is done and the data have been analyzed, the list will be destroyed. Your information will be combined with information from other people who took part in this study. When the researcher writes about this study to share what was learned with other researchers, He will write about this combined information. Your name will not be used in any report, so people will not know how you answered or what you did.

There are times when the researcher may have to show your information to other people. For example, the researcher may have to show your identity to people who check to be sure the research was done right. These may be people from the University of Central Florida or state, federal or local agencies.

Study contact for questions about the study or to report a problem:

Tareq Ahram, Graduate Student, Industrial Engineering & Mgmt. Systems , College of Engineering and Computer Science, (407) 823-0608 or by email at tahram@mail.ucf.edu or Dr. Pamela McCauley-Bush, Faculty Supervisor, Department of Industrial Engineering & Mgmt. Systems at (407) 823-6092, by email at mbush@mail.ucf.edu.

IRB contact about your rights in the study or to report a complaint: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). For information about the rights of people who take part in research, please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901.

If you are harmed because you take part in this study: If you believe you have been injured during participation in this research project, you may file a claim with UCF Environmental Health & Safety, Risk and Insurance Office, P.O. Box 163500, Orlando, FL 32816-3500 (407) 823-6300. The University of Central Florida is an agency of the State of Florida for purposes of sovereign immunity and the university's and the state's liability for personal injury or property damage is extremely limited under Florida law. Accordingly, the university's and the state's ability to compensate you for any personal injury or property damage suffered during this research project is very limited.

How to return this consent form to the researcher: Please write down your name and check all boxes that apply and return this consent form by email.

☐ I have read the procedure described above

☐ I voluntarily agree to take part in the procedure

☐ I am at least 18 years of age or older

_____ _____ _____

Signature of participant Printed name of participant Date

_____ _____

Principal Investigator Date

Study title:
The Multi-criteria Decision Weighted model to enhance information retrieval and search engines performance.

Study Description: Information retrieval today is much more challenging than traditional small document collection information retrieval systems. In this study, we focus on evaluating and testing a novel multi-criteria decision weighted model created to enhance information retrieval and search engines performance based on customized user priorities. Below is a brief description of each factor we would like to study:

- **Precision:** is the Proportion of relevant documents in the retrieved results to all returned results (relevant and non relevant).
- **Recall**: is the Proportion or relevant documents in the retrieved collection to the total number of relevant documents.
- **Relevance:** Documents relevancy (similarity) to search query in the retrieved results.
- **Search Length:** Expected position of a relevant document in returned results (How long are you willing to look into returned search result pages till you find relevant documents).
- **Query Processing time:** Time to process queries and return search results.

Based on your personal preferences and experience using web search engines (e.g. Google and Yahoo!). Please select the answer(s) which best match your priorities:

Please rate the relative importance of Precision with other factors:
1) Precision has_____ ***Recall***.

Absolutely less importance than — Equal importance to — Absolutely more importance than
○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

2) Precision has_____ ***Relevance***.

Absolutely less importance than — Equal importance to — Absolutely more importance than
○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

3) Precision has_____ ***Search length***.

Absolutely less importance than — Equal importance to — Absolutely more importance than
○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

4) Precision has_____ ***Processing time***.

Absolutely less importance than — Equal importance to — Absolutely more importance than
○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Please rate the relative importance of Recall with other factors:
5) Recall has_____ ***Relevance***.

Absolutely less importance than Equal importance to Absolutely more importance than

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

6) Recall has_____ ***Search Length***.

Absolutely less importance than Equal importance to Absolutely more importance than

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

7) Recall has_____ ***Processing time***.

Absolutely less importance than Equal importance to Absolutely more importance than

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Please rate the relative importance of Search length with other factors:

8) Search length has_____ ***Relevance***.

Absolutely less importance than Equal importance to Absolutely more importance than

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

9) Search length has_____ ***Processing Time***.

Absolutely less importance than Equal importance to Absolutely more importance than

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

Please rate the relative importance of Relevance with other factors:

10) Relevance has _____ ***Processing Time***.

Absolutely less importance than Equal importance to Absolutely more importance than

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Expedited Initial Review and Approval

From : **UCF Institutional Review Board**
FWA00000351, Exp. 6/24/11, IRB00001138

To : **Tareq Z Ahram**

Date : **September 08, 2008**

IRB Number: **SBE-08-05769**

Study Title: **Multicriteria Decision Weighted model to enhance information retrieval and search engines performance**

Dear Researcher:

Your research protocol noted above was approved by **expedited** review by the UCF IRB Vice-chair on 9/7/2008. **The expiration date is 9/6/2009.** Your study was determined to be minimal risk for human subjects and expeditable per federal regulations, 45 CFR 46.110. The category for which this study qualifies as expeditable research is as follows:

> 7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

The IRB has approved a **consent procedure which requires participants to sign consent forms.** Use of the approved, stamped consent document(s) is required. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Subjects or their representatives must receive a copy of the consent form(s).

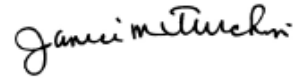
All data, which may include signed consent form documents, must be retained in a locked file cabinet for a minimum of three years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained on a password-protected computer if electronic information is used. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

To continue this research beyond the expiration date, a Continuing Review Form must be submitted 2 – 4 weeks prior to the expiration date. Advise the IRB if you receive a subpoena for the release of this information, or if a breach of confidentiality occurs. Also report any unanticipated problems or serious adverse events (within 5 working days). Do not make changes to the protocol methodology or consent form before obtaining IRB approval. Changes can be submitted for IRB review using the Addendum/Modification Request Form. An Addendum/Modification Request Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at http://iris.research.ucf.edu .

Failure to provide a continuing review report could lead to study suspension, a loss of funding and/or publication possibilities, or reporting of noncompliance to sponsors or funding agencies. The IRB maintains the authority under 45 CFR 46.110(e) to observe or have a third party observe the consent process and the research.

On behalf of Tracy Dietz, Ph.D., UCF IRB Chair, this letter is signed by:

Signature applied by Janice Turchin on 09/08/2008 03:51:45 PM EDT



IRB Coordinator

APPENDIX E: AHP ANALYSIS FOR SUBJECT MATTER EXPERT'S RESPONSES

Table: Subject Matter Experts Responses to questionnaire

Question#	SME#1	SME#2	SME#3	SME#4
1	3	4	6	6
2	4	3	4	4
3	2	4	4	4
4	4	5	7	5
5	7	4	3	4
6	6	5	5	4
7	4	3	5	3
8	6	5	2	5
9	7	5	5	4
10	7	7	6	4

Table: AHP analysis for SME performance measures ranking

Compare the relative importance with respect to: Goal: Matrix Intrinsic Dimensionality

	Precision	Recall	Relative Relevance	Average Search Length	Processing Time
Precision		1.0	3.0	3.0	1.0
Recall			1.0	1.0	3.0
Relative Relevance				1.0	1.0
Average Search Length					1.0
Processing Time	Incon: 0.08				

Table: AHP performance measures priorities

Priorities with respect to:
Goal: Matrix Intrinsic Dimensionality

Processing Time	.245
Relative Relevance	.235
Average Search Length	.235
Recall	.156
Precision	.128

Inconsistency = 0.08
with 0 missing judgments.

APPENDIX F: AVERAGE STANDARD ESTIMATOR RESULTS

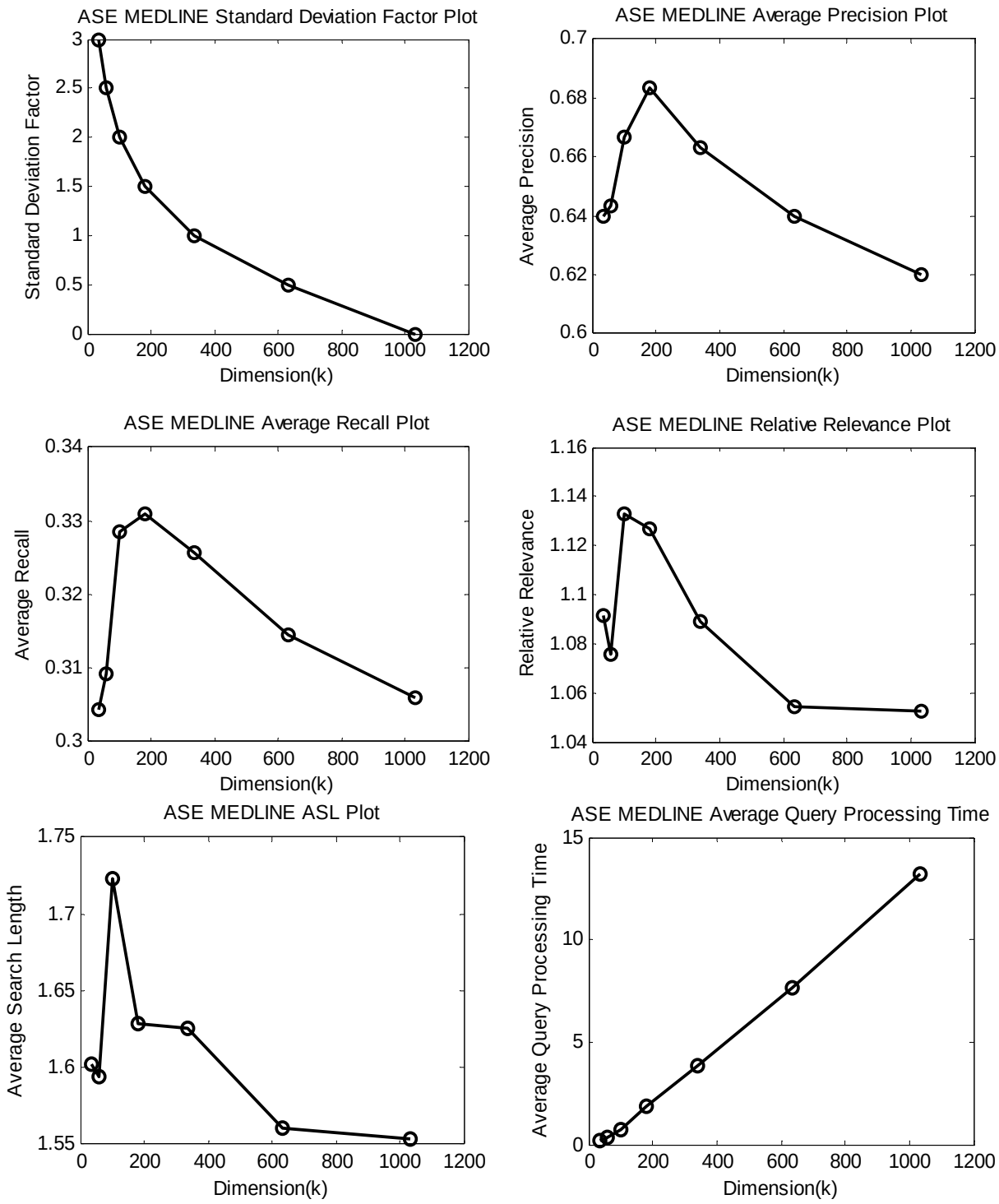


Figure: Performance measures plot for a range of dimensions using ASE standard deviation factor in MEDLINE

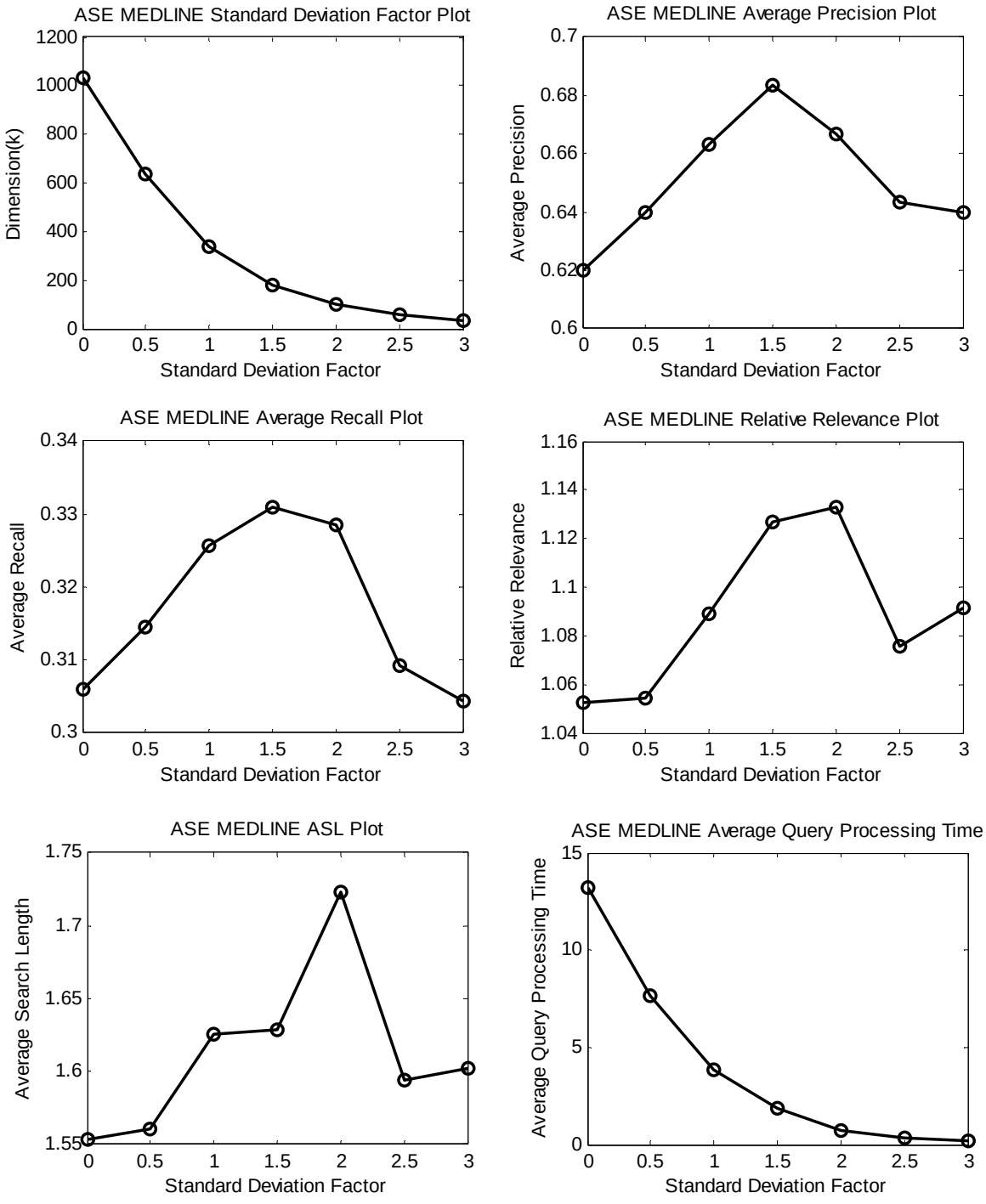


Figure: Performance measures plot for a range of ASE standard deviation factor in MEDLINE

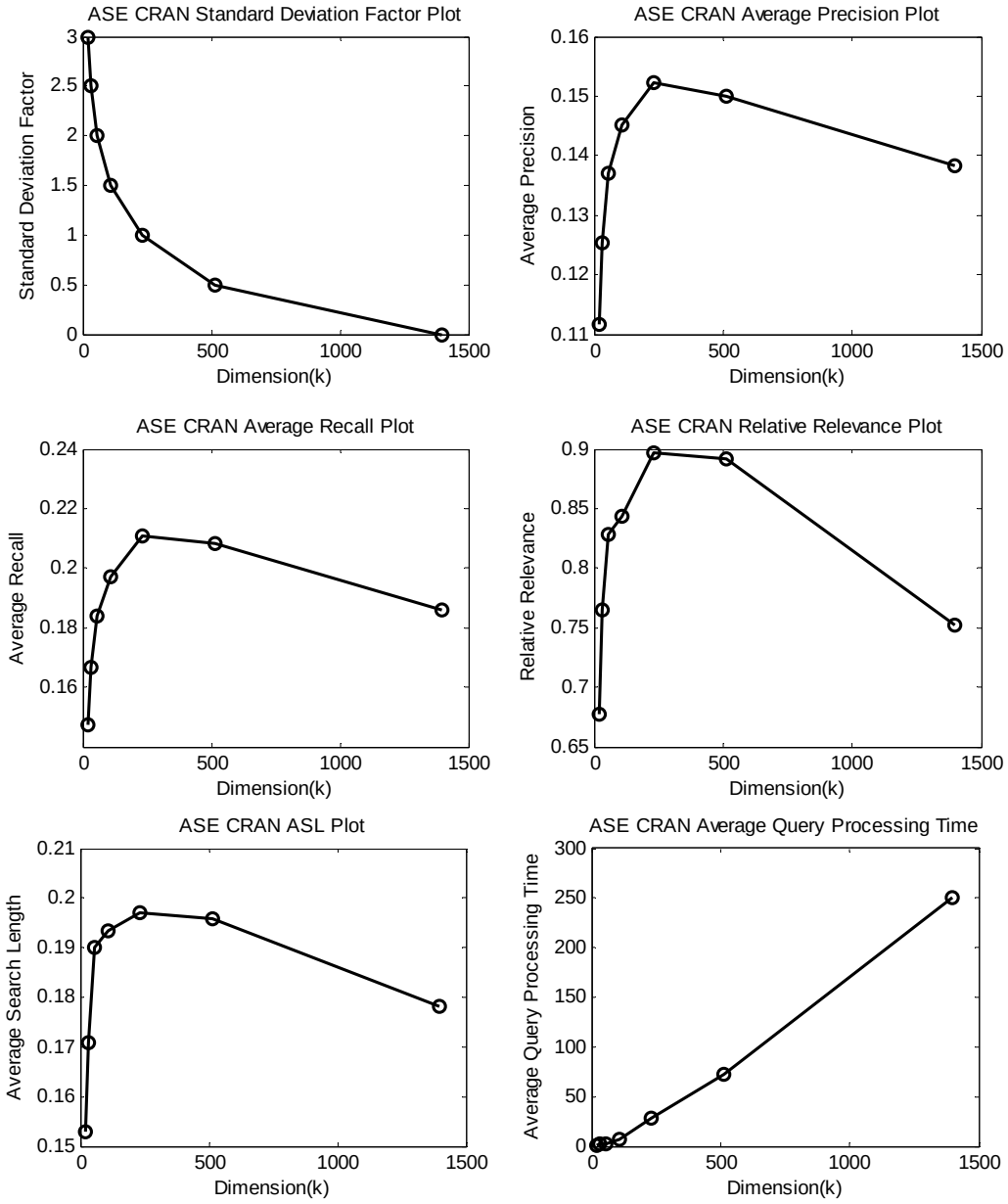


Figure: Performance measures plot for a range of dimensions using ASE standard deviation factor in CRANFIELD

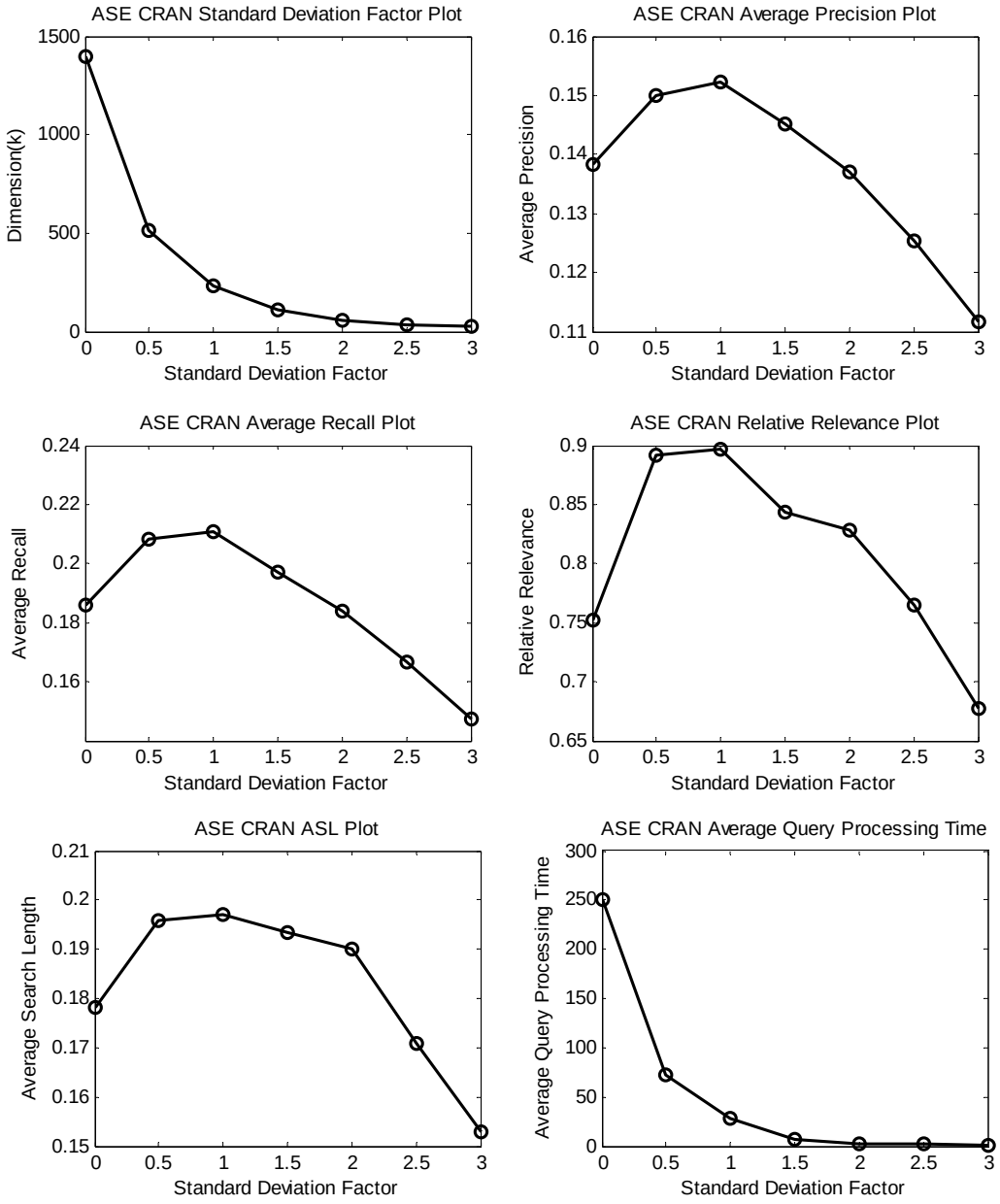


Figure: Performance measures plot for a range of dimensions using ASE standard deviation factor in CRANFIELD

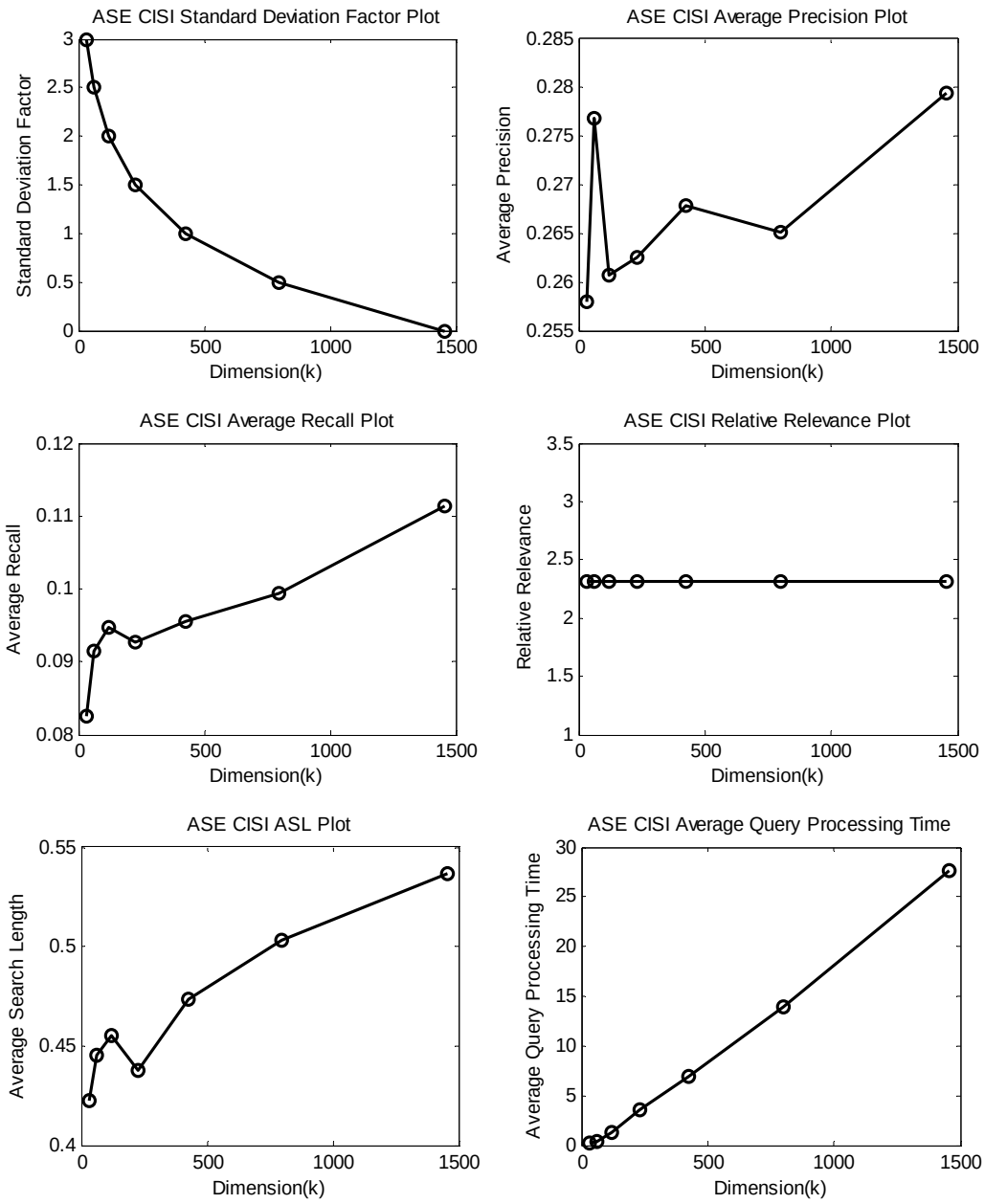


Figure: Performance measures plot for a range of dimensions using ASE standard deviation factor in CISI

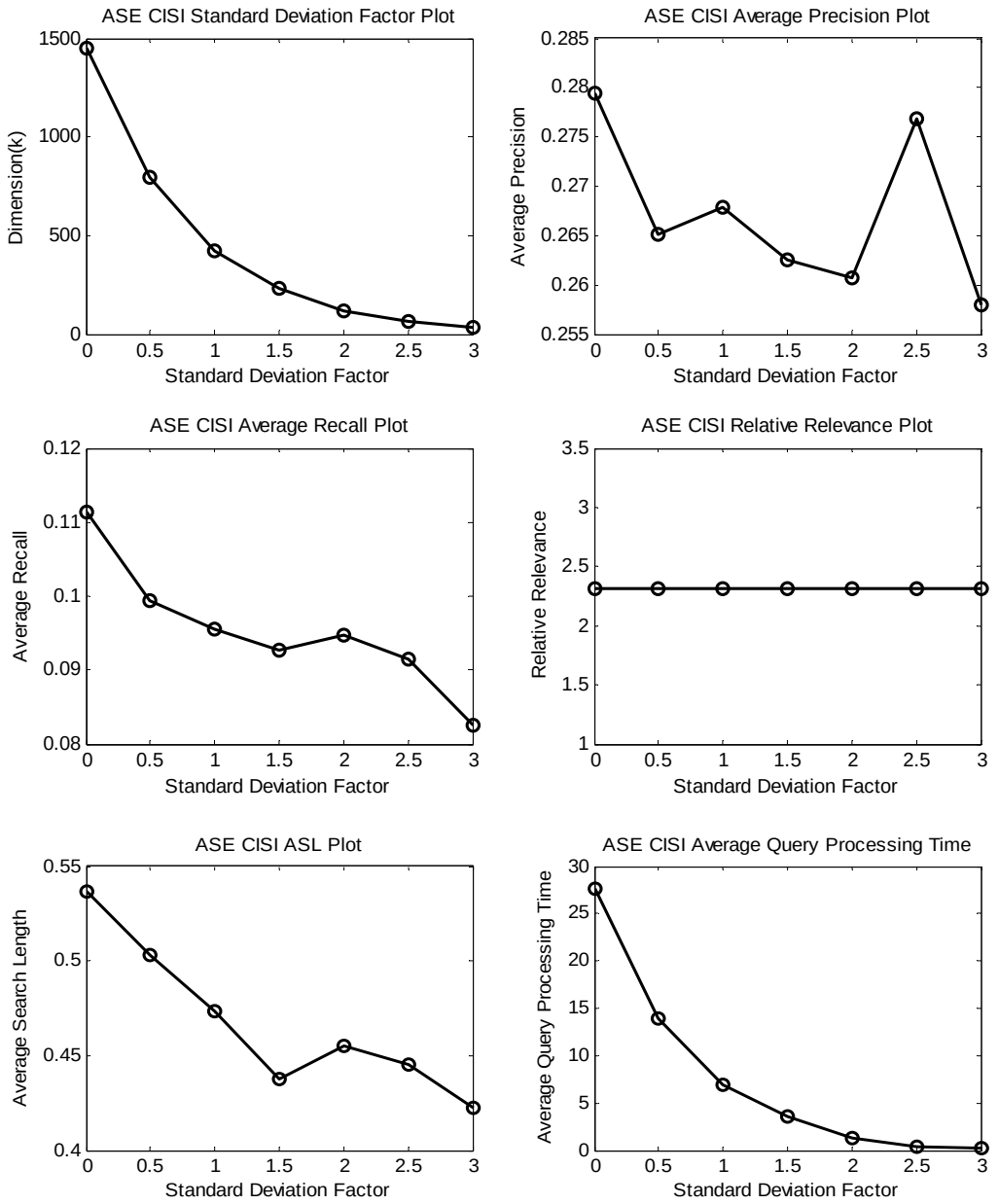


Figure: Performance measures plot for a range of dimensions using ASE standard deviation factor in CISI

LIST OF REFERENCES

1) Amsler, R. A. (1984). Machine readable dictionaries. In M. E Williams, editor, Annual Review of information Science and technology. Volume 19, pages 161-209. Knowledge Industry Publication, Inc.

2) Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J.Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, LAPACK User's Guide (http://www.netlib.org/lapack/lug/lapack_lug.html), Third Edition, SIAM, Philadelphia, 1999.

3) Anderson T. W. (1984). An introduction to multivariate statistical analysis. Wiley, 1984.

4) Back, Jonathan and Oppenheim, Charles (2001) "A model of cognitive load for IR: implications for user relevance feedback interaction". Information Research, 6(2)

5) Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval, Addison Wesley.

6) Burges, Christopher J.C.(2004) Geometric Methods for Feature Extraction and Dimensional Reduction, Technical Report MSR-TR-2004-55 Microsoft Research

7) Beaulieu, M. (1997). Experiments on interfaces to support query expansion. Journal of Documentation, 53, (1), 8-19.

8) Berry, M. and Browne, M. (1999). Understanding Search Engines: Mathematical Modeling and Text Retrieval, Society for Industrial and Applied Mathematics.

9) Berry, M.W. , Dumais S. T & G.W (1994). Using Linear Algebra for Intelligent Information Retrieval. O'Brien Computer Science Department CS-94-270.

10) Bhattacharyya, G. K., and Johnson R. A. (1977). Statistical Concepts and Methods. Wiley.

11) Bollacker, K., Lawrence, S. and Giles C. Lee. (1998) CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. In Sycara and M. Wooldridge, editors, Proceeding of the Second International Conference on Autonomous Agents, PAGES 116-123, New York, ACM Press

12) Bookstein, A. and Swanson D. R. (1975). A Decision Theoretic Foundation for Indexing. Journal of the American Society for Information Science, 26(1):45-50.

13) Borlund, P. & Ingwersen, P. (1998) Measures of relative relevance and ranked half-life: performance indicators for interactive IR. In: Croft, B.W, Moffat, A., van Rijsbergen, C.J., Wilkinson, R., and Zobel, J., eds.

14) Borg, I.and Groenen.P. (1997) Modern multidimensional scaling: theory and applications- Modern multidimensional scaling: theory and applications,Springer.

15) Borlund, Pia (2003) "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems"Information Research, 8(3), paper no. 15

16) Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems 30(1-7): 107-117. From URL: http://infolab.stanford.edu

17) Burnham K. P. and Anderson D. R. (1998). Model selection and inference: A practical Information Theoretic approach, Springer, New York, 1998

18) Cherkassky, V. S., Mulier, F. (1998) Learning from Data: Concepts, Theory, and Methods - John Wiley & Sons, Inc. New York, NY, USA

19) Chowdhury, G. G. (1999). Introduction to Modern Information Retrieval, Library Association Publishing.

20) Church, K. W., Hanks, P.,(1990). Word association norms, mutual information, and lexicography - Computational Linguistics.

21) Cleverdon C. W.(1967). The cranfield tests on index language devices. ASLIB Proceedings, 19:173192. 1967.

22) Cleverdon, C. W., Mills, J.(1963). The testing of index language devices. ASLIB Proceedings,15:106-130

23) Computer Industry Almanac Inc., (2003), Internet Industry Almanac. , from URL: http://www.c-i-a.com/internetuseres.htm

24) Cooper W. S. (1988) Getting beyond Boole. Information Processing and Management, 24:243-248.

25) Cooper, W. S. (1991). Some Inconsistencies and Misnomers in Probabilistic Information Retrieval. In proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 57-62.

26) Cooper, W. S. and Maron, M. E. (1978). Foundation of probabilistic and utility theoretic indexing. Journal of the ACM, 25(1):67-80.

27) Cooper, WS. (1973). On selecting a measure of retrieval effectiveness. Part I.

28) Cooper, G. (1998). Research into cognitive load theory and instructional design at UNSW. University of New South Wales, Sydney, Australia.

29) Cosijn, E. & Ingwersen, P. (2000). Dimensions of relevance. Information Processing & Management, 36(4) 533-550.

30) Cox, T. F. and Cox, M. A. (2001).Multidimensional Scaling number 88 in Monographs on Statistics and Probability Wiley 2nd Ed.

31) Croft W. B. (1987). Approaches to intelligent information retrieval Information Processing and Management, 23(4):249254.

32) Croft, W. B. (1986). User specified domain knowledge for document retrieval In ACM Conference on Research and Development in Information Retrieval, pages 201-206. Association for Computing Machinery.

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

34) Dillon, W. and Goldstein, M. (1984). Multivariate Analysis: Methods and Applications. Wiley.

35) Ding. C. H. Q. (1999). A similarity based probability model for latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

36) Ding. C. H. Q. (2000). A probabilistic model for dimensionality reduction in information retrieval and filtering. Read at 1st SIAM Computational Information Retrieval Workshop.

37) Dodgson, M. (2000). The management of technological innovation an international and strategic approach. Oxford [England]: Oxford University Press.

38) Dumais S. T. (1992) LSI meets TREC: A status report." In: D. Harman (Ed.), The First Text Retrieval Conference (TREC1), National Institute of Standards and Technology Special Publication 500-207, pp. 137-152.

39) Dumais S. T. (1993) Latent semantic indexing (LSI) and TREC2. In Proceedings of the Second Text Retrieval Conference (TREC 2).

40) Dumais S. T. (1994) Latent semantic indexing (LSI): TREC3 report. In Proceedings of the Third Text Retrieval Conference (TREC 3).

41) Efron, B. (1993). An Introduction to the Bootstrap. Chapman and Hall.

42) Efron, M. (2003). Eigenvalue-based Estimation for optimal dimensionality reduction in information retrieval. Dissertation, Chapel Hill.

43) Erkut, E., Moran, SR. (1991) Locating obnoxious facilities in the public sector: an application of the hierarchy process – Socio-Economic Planning Sciences.

44) Fisher, R. A. (1974). Collected papers of R. A. Fisher (1971-1974). University of Adelaide.

45) Fodor I. K. (2002). "A survey of dimension reduction techniques," LLNL technical report, June 2002, UCRL-ID-148494

46) Foltz, P., Kintsch, W., and Landauer T. (1998). The measurement of textual coherence with latent semantic analysis. Discourse Processes, 25(2&3); 285-307.

47) Forsythe, G., Malcom, M., and Moler, C. (1977) Computer methods for mathematical Computations (ch-9). Prentice Hall.

48) Fox, E. A. (1980). Lexical relations; Enhancing effectiveness of information retrieval systems.ACM SIGIR Forum, 15(3); 536.

49) Frakes, B. (1992). Stemming algorithms, in B. Frakes and R. Baeza-Yates (eds), Information Retrieval Data Structures and Algorithms, Morgan Kaufmann, San Francisco, CA, pp. 131-160.

50) Frakes, W. B. and Baeza-Yates, R. (1992). Information Retrieval: Data Structures and algorithms, Prentice Hall.

51) Fukunaga, K. and Olsen, D.R. (1971). An algorithm for finding intrinsic dimensionality of data. IEEE Transactions on Computers, C-20:176–183.

52) Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human system communication. Communications of the ACJM, 30(11); 964-971.

53) Gardenfors, P. (2000) Conceptual Spaces: The Geometry of Thought. MIT Press.

54) Godwin G. Udo. (2000) Using analytic hierarchy process to analyze the information technology outsourcing decision. Industrial Management & Data Systems 100/9 [2000], p.p 421-429.

55) Golub, G. H. and Van Loan, C. F. (1989). Matrix Computations. Baltimore, Johns Hopkins University Press.

56) Gulli, A.and Signorini, A.. (2005). Building an open source meta search engine. In 14th WWW, from URL: http://www.cs.uiowa.edu

57) Guttman, L. (1954). Some necessary conditions for common factor analysis. Psychometrika, 19(2):149-161.

58) Harman, D. K., ED.(1995). The 3rd Text Retrieval Conference (TREC-3). NIST Special Publication 500-225. http://trec.nist.gov

59) Harter S. P., Hert C. A. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. In M. E. Williams, editor, Annual Review of Information Science and Technology, volume 32, pages 394. American Society for Information Science.

60) Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, New York.

61) Hofmann T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42:177-196.

62) Hofmann T. (1999). Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, pages 50-57, Berkeley, California.

63) Hofmann T., (2000). Learning probabilistic models of the web. In Research and Development in Information Retrieval, pages 369-371.

64) Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika. 30:179-186.

65) Hu, P.J.H., Ma, P.C., & Chau, P.Y.K. (1999). Evaluation of user interface designs for information retrieval systems: a computer-based experiment. Decision Support Systems, 27, (1-2), 125-143.

66) Huang, L. (2000). A survey on web information retrieval technologies.

67) Husbands, P., Simon, H., and Ding, C. (2000). The use of singular value decomposition for text retrieval.

68) Hutchins, W. J. (1978). The concept of "Aboutness" in subject indexing. Aslib Proceedings, 30, 172-181

69) Huurnink B. (2005). Toward a fully automated video search engines Thesis, University of Amsterdam

70) Ide, E. (1971). New experiments in relevance feedback. In G. Salton. Editor. The SMART Retrieval System, pages 337354. Prentice Hall.

71) Jackson, J. E. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. Ecology, 74:2204-2214.

72) Jain, A.K. Wyse N., Dubes R.(1980). A critical evaluation of intrinsic dimensionality algorithms. In E.S. Gelsema and 1.N. Kanal. editors, Pattern Recognition in Practice, pages 415-425. North-Holland.

73) Jansen, B. J. (2000). An investigation into the use of simple queries on Web IR systems. Information Research. : An Electronic Journal. 6(1).

74) Jiang F. and Littman M. L. (2000). Approximate dimension equalization in vector based information retrieval In Proc. 17th International Conf. on Machine Learning, pages 423-430. Morgan Kaufmann, San Francisco, CA.

75) Joachirns T. (1998). Text categorization with support vector machines: learning with many relevant features. In Claire Nedellec and Celine Rouveirol, editors, Proceedings of ECML-98, 10th European Conference on machine Learning, pages 137-142, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

76) Jobson J. D.(1991). Applied multivariate Data Analysis. Springer.

77) Jolliffe L T. (2002) Principal Component Analysis. Springer, 2nd edition.

78) June, Wei., (2004), Global Competitive Internet Usage Forecasting Across Countries and Languages, EDSIG 2004

79) Jurafsky, D. and Martin, J. (2000). Speech and Language Processing, Prentice Hall.

80) Jurafsky, D. and Martin, J. H. (1999). Speech and Language Processing. Prentice Hall.

81) Kawasaki, Y.; Sunahara, H. (2000**).** An architecture of the distributed multimedia information retrieval network with query routing systems, Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on Volume 2, Issue , Page(s):1179 - 1182 vol.2

82) Kise, K., Junker, M., Dengel, A. and Matsumoto K. (2001). Experimental evaluation of passage-based document retrieval, Proceedings of the 6th International Conference on Document Analysis and Recognition, pp. 592-596.

83) Kohonen, Teuvo (2001). Self-organizing maps. 3rd ed. Berlin: Springer.

84) Kolda, T. (1997). Limited-Memory Matrix Methods with Applications, PhD thesis, University of Maryland at College Park, Applied Mathematics Program. URL: citeseer.nj.nec.com/115586.html

85) Kolda, T. G. and O'Leary, D. P. (1998). A semi-discrete matrix decomposition for latent semantic indexing information retrieval, ACM Transactions on Information Systems 16(4): 322-346.

86) Kolda, T. G. and O'Leary, D. P. (2000). Algorithm 805: Computation and uses of the semi discrete matrix decomposition, ACM Transactions on Mathematical Software 26(3): 415-435. URL: http://doi.acm.org/10.1145/358407.358424

87) Krzanowski, W. J. (1988) Principles of Multivariate Analysis: A User's Perspective. Oxford University Press.

88) Kuhlthau, C. (1993). A principle of uncertainty for information seeking. Journal of Documentation, 49, 4, 339-355.

89) Kuhlthau, C., Spink, A., & Cool, C. (1992). Exploration into stages in the information search process in online information retrieval. Proceedings of the ASIS annual meeting, 29, 67-71.

90) Landauer T. K. and Dumais S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2):211-240.

91) Landauer, TK., Foltz, PW., Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284

92) Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. The Psychology of Learning, 41:43-84.

93) Landauer, T. K., Laham, D., and Foltz, P. (1998). Learning humanlike knowledge by singular value decomposition: A progress report. In M. L Jordan, M. J. Kearns, and S. A. Solla. Editors. Advances in Nev.ml Information Processing Systems, volume 10. The MIT Press.

94) Lawrence, S. & Giles, C. (1999). Accessibility of information on the web. Nature, 400, 107 – 109.

95) Laurence, S. & Margolis, E. (1999). Concepts and cognitive science. In Concepts: Core Readings, pages 382. MIT Press.

96) Lesk, M. E. (1969). Word word associations in document retrieval systems. American Documentation, 20 (1): 27-38.

97) Levina, E. and Bickel, P.J.(2004). Maximum likelihood estimation of intrinsic dimension. In Advances in Neural Information Processing Systems, volume 17, Cambridge, MA, USA, 2004. The MIT Press.

98) Liddy, E. (2001). How a search engine works. From URL: http://www.infotoday.com/searcher/may01/liddy.htm

99) Lootsma, F. A. (1999). Multi-criteria decision analysis via ratio and difference judgement. Applied optimization, v. 29. Boston: Kluwer.

100) Losee, R. M. (1994). Term dependence: Truncating the bahadur lazarsfeld expansion. Information Processing and Management, 30(2):293-303.

101) Losee, R. M. (1998). Text Retrieval and Filtering: Analytic Models of Performance. Kluwer, Boston.

102) Losee, R. M. (2000). When information retrieval measures agree about the relative quality of document rankings. Journal of the American Society for Information Science,51(9):834-840.

103) Luhn H. P. (1961) The automatic derivation of information retrieval encodements from machine readable texts. In A. Kent, editor, Information Retrieval and Machine Translation, volume 3, pages 1021-1028. Interscience Publication.

104) Luhn, H. P. (1955). A new method of recording and searching information. American Documentation,4(1):14-16.

105) Luhn, H. P. (1957). A statistical approach to the mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1(4):309-317, October 1957.

106) Maaten Laurens van der (2007). An Introduction to Dimensionality Reduction Using Matlab (Report MICC), Maastricht University.

107) Maaten Laurens van der, Postma E.O., and Herik. H.J. van den (2007). Dimensionality reduction: A comparative review. IEEE Transactions on Pattern Analysis and Machine Intelligence.

108) Malczewski, J. Moreno-Sanchez, R. Bojorquez-Tapia LA. (1997) Multicriteria Group Decision-making Model for Environmental Conflict Analysis in the Cape Region, Mexico. Journal of Environmental Planning and Management. Volume 40, Number 3, pp. 349-374(26)

109) Malczewski, J. (1999) GIS and Multicriteria Decision Analysis. John Wiley and Sons

110) Manning, C. and Schutze, H. (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge.

111) Manning, C. D. and SchÄutze, H. (1999). Foundations of statistical natural language processing, MIT Press.

112) McCullagh, P. and Nelder, J. (1989). A. Generalized Linear- Models. Chapman and Hall, Boca Raton. second edition.

113) Mihail, M. and Papadimitriou, C. H. (2002). On the eigenvalue power law. Read at RANDOM.

114) Miller, G.A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

115) Muknahallipatna, S. and Chowdhury, B.H. (1996). Input dimension reduction in neural network training-case study intransient stability assessment of large systems Dept. of Electr. Eng., Wyoming Univ., Laramie, WY.

116) Neter, J., Kutner, M. H.,. Nachtsheim, C. J, and Wasserman W. (1996). Applied Linear- Statistical Models. Irwin, Chicago.

117) Newby, G. B. (2001) Cognitive space and information space. Journal of the American Society for Information Science, 52(12):10261048.

118) Nielsen//NetRatings MegaView Search, April (2006), from URL: http://www.nielsen-netratings.com/

119) Nielsen/NetRating (2000). Retrieved from the World Wide Web on 11 August 1999, from URL: http://www.nielsen-netratings.com/.

120) Nielsen, Jakob. (1993). Usability Engineering. Academic Press, Boston, MA.

121) Nielsen, Jakob.(1997) The need for speed. Alertbox (web page: http://www.useit.com/alertbox/9703a.html).

122) Oakes. M. P. (1998). Statistics for- Corpus Linguistics. Edinburgh University Press, Edinburgh.

123) Papadimitriou, C. H., Tamaki, H., Raghavan, P., and Vempala S.(1998). Latent semantic indexing: A probabilistic analysis. In Proceedings of the Seventeenth ACM SIGMOD-SIGART Symposium on Principles of Database Systems, pages 159-168.

124) Porter, M. F. (1980). An algorithm for suffix stripping. In Program, pages 130-137.

125) Prey, K., French, J. C., Powell, A. 1., and Viles, C. L.(2001). Inverse document frequency and web search engines. Technical Report CS-2001-07, University of Virginia.

126) Quine, W. V. O. (1999). Two dogmas of empiricism. In Concepts: Core Readings, pages 153 170. MIT Press.

127) Raymer, M.L. Punch, W.F. Goodman, E.D. Kuhn, L.A. Jain, A.K. (2000) Dimensionality reduction using genetic algorithms. Dept. of Biol., Michigan State Univ., East Lansing, MI;

128) Rencher, A. C. (1995). Methods of Multivariate Analysis. Wiley Interscience.

129) Rocchio, J. J. (1971). Relevance feedback in information retrieval In G. Salton. Editor. The SMART Retrieval System Experiments in Automatic Document Processing. Prentice Hall.

130) Rosch, E. (1975). Family resemblance: Studies in the internal structure of categories. Cognitive Psychology, 7:573-605.

131) Rosch, E. (1999) Principles of categorization. In Concepts: Core Readings, pages 189206. MIT Press.

132) Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., and Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8:382-349.

133) Saaty, TL. (1980).The Analytical Hierarchy Process- McCraw-Hill New York.

134) Saaty, TL. (1990). Multicriteria Decision Making: The Analytic Hierarchy Process: Planning, Priority Setting, Resource…, RWS Publications.

135) Salton G. and McGilL M. (1983). Introduction to Modern Information Retrieval. McGraw Hill.

136) Salton G. and McGill, M. (1983) Introduction to Modern Information Retrieval. McGraw Hill.

137) Salton G., Yang C. S., and Yu C. T. (1975). A theory of term importance in automatic text analysis. Journal of the American Society for Information Science, 26(1):3344.

138) Salton, G. (1989) Automatic Text Processing. Addison-Wesley.

139) Salton, G. (1989). Automatic text processing: the transformation, analysis, and retrieval of information by computer, Addison-Wesley.

140) Salton, G. and Buckley C. (1988). Term weighting approaches in automatic text retrieval Information Processing and Management, 24:513-523, 1988.

141) Salton, G. and Buckley, C. (1997b). Term-weighting approaches in automatic text retrieval, in K. Sparck Jones and P. Willet (eds), Readings in Information Retrieval, Morgan Kaufmann Publishers, Inc.

142) Salton, G. and Lesk, M. E.(1968). Computer evaluation of indexing and text processing. Journal of the ACM, 15(1):836.

143) Salton, G. and McGill, M. J.(1983). The SMART and SIRE Experimental Retrieval System. McGraw Hill.

144) Salton, G., Wong, A., and Yang, C. S. (1975) A vector space model for automatic indexing. Communications of the ACM, 18:613-620.

145) Saracevie, T. (1975). Relevance: A review of and a framework for thinking on the notion in information science. Journal of the American Society for Information Science, 26:321-343.

146) Saracevic, T. (1996) Relevance reconsidered '96. In: Ingwersen, P. & Pors, N.O., eds. Proceedings of CoLIS 2, Second International Conference on Conceptions of Library and Information Science: Integration in Perspective. Copenhagen 1996. Copenhagen: Royal School of Librarianship, pp. 201-218.

147) Schamber, L. (1994). Relevance and information behavior. In M. E. Williams. Editor. Annual Review of Information Science and Technology, volume 29, pages 3-48. American Society for Information Science.

148) Shaw, W. M., Burgin, R. and Howell, P. (1997). Performance standards and evaluation in test collections: Cluster based retrieval models. Information Processing and Management, 33(1):1-14

149) Sparck Jones K. (1972). A statistical interpretation of term specificity and its application in retrieval Journal of Documentation, 28(1):11-21.

150) Sparck Jones K. (1979). Search term relevance weighting given little relevance information. Journal of Documentation, 35:30-48.

151) Sparck Jones, K. and Tail, K. (1984). Automatic search term variant generation. Journal of Documentation, 40(1):50-66.

152) Sperber, D. and Wilson, D.(1995). Relevance: Communication and Cognition. Blackwell, 2nd Ed.

153) Squire, David McG. Muller, Henning., Muller Wolfgang, (1999) "Improving Response Time by Search Pruning in a Content-Based Image Retrieval System, Using Inverted File Techniques," p. 45, IEEE Workshop on Content-Based Access of Image and Video Libraries.

154) Story R. E. (1996). An explanation of the effectiveness of latent semantic indexing by means of a Bayesian regression model Information Processing and Management, 32(3):329-344.

155) Strang, G. (1998). Linear- Algebra and its Applications. International Thompson Publishing.

156) Subhash, S. (1996). Applied Multivariate Techniques. Wiley.

157) Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257-285.

158) Tang B., Shepherd M., Heywood MI., Luo X. (2005) Comparing Dimension Reduction Techniques for Document Clustering. Advances in Artificial Intelligence Volume 3501/2005, Springer Link.

159) Van Rijsbergen C. J. (1979). Information Retrieval. Butterworths, 2nd edition.

160) Van Rijsbergen, C. H. (1977). A theoretical basis for the use of co-occurrence data in information retrieval Journal of Documentation, 33(2):106-119.

161) Voorhees, Ellen M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the ACM SICIR Conference on Research and Development in Information Retrieval, pages 315-323.

162) Wang, Ye Diana. , Forgionne, Guisseppi. , (2005). A decision-theoretic approach to the evaluation of information retrieval systems, Information Processing and Management: an International Journal Volume 42 , Issue 4

163) Wittgenstein, L. (1953). Philosophical Investigations. Prentice Hall, 3rd edition.

164) Wong, S. K. M., Ziarko W., Raghavan, V. V. and Wong, P. C. N. (1987) On modeling of information retrieval concepts in vector space. TODS, 12(2):299-321.

165) Xiao Luo; Zincir-Heywood, A.N. (2004). Evaluation of three dimensionality reduction techniques for document classification, Electrical and Computer Engineering, 2004. Canadian Conference on Volume 1, Issue, 2-5 May 2004 Page(s): 181 - 184 Vol.1.

166) Xu, J. L. (1999) Internet Search Engines: Real World IR Issues and Challenges. Paper presented at the Conference on Information and Knowledge Management. Kansas City, Missouri.

167) Zanakis, SH. , Solomon, A. , Wishart, N., Dublish, S. (1998). Multi-attribute decision making: A simulation comparison of select methods, European Journal of Operational Research, Elsevier

168) Zeimpekis, D., Gallopoulos, E. (2007) Text to Matrix Generator_User's Guide, Department of Computer Engineering and Informatics, University of Patras, Greece.

169) Zipf, G. K.(1929). Relative frequency as a determinant of phonetic change. Harvard Studies in Classical Philology, 40:195.