Latent Semantic Analysis: Five methodological recommendations

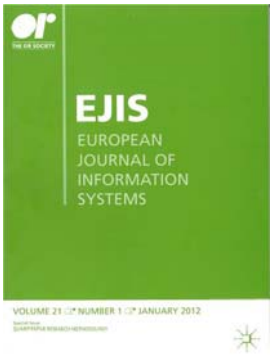
Nicholas Evangelopoulos,
Information Technology and Decision Sciences Department,
College of Business, University of North Texas,
P.O. Box 305249 - BUSI 336,
Denton, Texas 76203, USA.
Tel: +1-(940) 565-3056
Fax: +1-(940) 565-4935
E-mail: nick.evangelopoulos@unt.edu

Xiaoni Zhang,
Department of Business Informatics,
College of Informatics, Northern Kentucky University,
Highland Heights, Kentucky, USA.

Victor R. Prybutok,
Information Technology and Decision Sciences Department,
College of Business, University of North Texas,
Denton, Texas, USA.

Cite as:

Evangelopoulos, N., Zhang, X., and Prybutok, V. (2012), "Latent Semantic Analysis: Five Methodological Recommendations." *European Journal of Information Systems*, **21**(1), January 2012 [Special Issue on Quantitative Methodology], pp. 70-86. DOI: 10.1057/ejis.2010.61. Published online 21 December 2010.



About the authors

Nicholas Evangelopoulos is an Associate Professor of Decision Sciences at the University of North Texas and a Fellow of the Texas Center for Digital Knowledge. His research interests include Statistics and Text Mining. His publications include articles appearing in *MIS Quarterly*, *Communications in Statistics*, and *Computational Statistics & Data Analysis*.

Xiaoni Zhang is an Associate Professor of Business Informatics at the Northern Kentucky University. She received her Ph.D. in Business Computer Information Systems from the University of North Texas in 2001. Her publications appear in *IEEE Transactions on Engineering Management*, *Communications of the ACM*, *International Conference of Information Systems* and *Information & Management*.

Victor R. Prybutok is a Regents Professor in the Information Technology and Decision Sciences Department in the College of Business and the Associate Dean of the Toulouse Graduate School at the University of North Texas. Dr. Prybutok is an ASQ certified quality engineer, certified quality auditor, and certified quality manager. Dr. Prybutok has authored over 90 journal articles and more than 70 conference presentations.

Latent Semantic Analysis: Five methodological recommendations

Abstract

The recent influx in generation, storage and availability of textual data presents researchers with the challenge of developing suitable methods for their analysis. Latent Semantic Analysis (LSA), a member of a family of methodological approaches that offers an opportunity to address this gap by describing the semantic content in textual data as a set of vectors, was pioneered by researchers in psychology, information retrieval, and bibliometrics. LSA involves a matrix operation called singular value decomposition, an extension of principal component analysis. LSA generates latent semantic dimensions that are either interpreted, if the researcher's primary interest lies with the understanding of the thematic structure in the textual data, or used for purposes of clustering, categorisation and predictive modelling, if the interest lies with the conversion of raw text into numerical data, as a precursor to subsequent analysis. This paper reviews five methodological issues that need to be addressed by the researcher who will embark on LSA. We examine the dilemmas, present the choices, and discuss the considerations under which good methodological decisions are made. We illustrate these issues with the help of four small studies, involving the analysis of abstracts for papers published in the *European Journal of Information Systems*.

Keywords: text mining, analysis of textual data, singular value decomposition, clustering, factor analysis.

Introduction

Textual data appear in an ever-increasing number of business and research situations, and are encountered by Information Systems (IS) researchers in a variety of contexts. For example, researchers in the IS discipline have an interest in examining titles, abstracts, or full-text bodies of IS publications in order to identify attributes such as research topics, theories, and methods, related to the nature of the research (Larsen et al. 2008, Sidorova et al. 2008, Willcocks et al. 2008, Dwivedi and Kuljis 2008, Hovorka et al. 2009). While such examination of textual data is frequently done qualitatively by the researcher who will apply expert judgement, the growing amount of textual data suggests a value in utilising a quantitative method, especially when the researcher opts for a more inclusive selection of journal sources (Larsen and Monarchi 2004, Larsen et al. 2008). Another example involves strategic and organisational IS researchers who study IS in their business, political, and societal environments. Such researchers often examine large volumes of corporate announcements, regulatory body statements, or corporate Web documents, in order to identify content attributes that can be related to organisational or social phenomena (Meroño-Cerdan and Soto-Acosta 2007, Spomer 2009). Once again, the amount of textual data may discourage the researcher from manual qualitative examination. The list of contexts in which the IS researcher may encounter textual data includes a large number of additional applications, such as IS development researchers who examine system requirement documents in order to propose efficient methods for translating them into formal designs, (Bajwa et al. 2009) as well as E-Commerce researchers who study descriptions of EDI standards in an effort to identify attributes in the technical language (Damsgaard and Truex 2000) or the implied points-of-view of the key participants (Barrett 1999) that can be related to EDI adoption. In these

application domains, IS researchers are potentially interested in research questions that fall under the following broad categories:

(1) which attributes, in the form of pre-defined categories, naturally emerging inherent categories, or latent semantic factors, are relevant to structuring the body of textual data?

(2) which attributes of textual data are related to particular outputs of the information system under study or its social and business environment?

Examples of more specific research questions would include:

(3) how homogeneous are the semantic factors in systems requirements over various communities involved in requirements gathering?

(4) what factors describe the semantic content in EDI adoption?

Traditionally these questions are addressed by associating textual data units to *a priori* proposed attributes through manual or automated content analysis. Content analysis is defined as a systematic, replicable technique for reducing a large body of text into content categories based on explicit rules of coding (Weber 1990). Content analysis offers a bridge between textual data and quantitative analysis and has been employed in IS research extensively, especially in the analysis of interviews (Dam and Kaufmann 2008), open-ended surveys (Couger and O'Callagher 1994, Panteli et al. 1999) or customer feedback (Ghose 2009). Traditional content analysis is cited as having the potential to contaminate coded output by the theoretical prejudices of the researcher who compiles the coding protocol (Franzosi 2004, p. 60). In this article we provide some methodological guidance on the appropriate use of Latent Semantic Analysis (LSA), a member of a family of quantitative methods that lie at the intersection of automated content analysis and information retrieval and provide for more objective approaches to the analysis of textual data used by researchers in order to answer research questions such as the ones listed

above. Going beyond simple key word discovery LSA describes the semantic content in textual data as a set of vectors and provides the opportunity to glean insights from the text that was not predicated upon a set of a priori assumptions, such as a predefined list of key words, because it provides a methodological approach to determining the categories. Similar to what is done in content analysis, LSA is a methodology that provides input into post-LSA procedures which allow for coding textual data into categories but it can also serve as a methodological aid in knowledge acquisition and retrieval. The potential benefits from employing LSA include (1) avoiding human subjectivity when the categories are pre-existing and (2) distilling new, data-driven categories when there is absence of well-established theories that *anticipate* the coding categories. The analytic approaches associated with LSA include numerically comparing documents to one another, developing new categories from a collection of documents and classifying documents into pre-existing categories and summarising a collection of documents with the help of a small number of interpretable dimensions. To better illustrate the methodological issues we provide four small studies involving the analysis of abstracts for papers published in the *European Journal of Information Systems*. The main focus of these studies is the text summarisation applications of LSA that are little researched and have application potential in IS research. However, the methodological considerations we address are also applicable to a broader analytic scope that includes information retrieval, document comparisons, document categorisation and quantification of textual data as a precursor to predictive modelling. Our paper is organised as follows. The next section is a brief introduction to LSA. Subsequent sections discuss some important methodological considerations related to various analysis stages of LSA that include the type of quantitative analysis, term filtering, term

weighting, dimensionality reduction and threshold selection. The paper concludes with a summary of our recommendations.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) originated in the late 1980s (Deerwester et al. 1990) as an information retrieval technique designed to improve library indexing and search engine query performance (Dumais 2004, Dumais 2007, Manning et al. 2008 pp 369-384). It was later proposed by psychology researchers as a theory and method for extracting and representing the meaning of words by humans, including word sorting and category judgments (Landauer 2007). The fundamental idea behind LSA is that the meaning of each passage of text (a *document*) is related to patterns of presence or absence of individual words, whereas a collection of documents (a *corpus*) is modelled as a system of simultaneous equations that can determine the similarity of meaning of words and documents to each other. A truncated representation of the original structure was shown to drastically improve query performance in part because it reduces the adverse effects of synonymy and polysemy.

Research interest on LSA spans the fields of information retrieval, artificial intelligence, psychology, cognitive science, education, information systems, and many others. Since the early 2000s, as the availability of stored text exploded, LSA experienced an explosion in popularity. A query for *Latent Semantic Analysis* or *Latent Semantic Indexing* on multiple electronic library databases yielded a total of 259 research journal articles published in the 1989-2009 period as shown in Figure 1. However, in spite of its application potential, LSA has received little attention

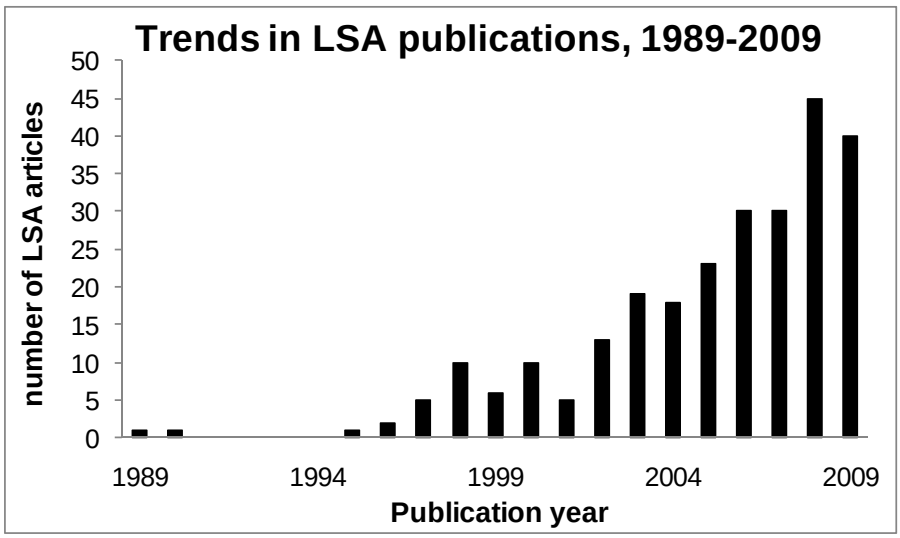


Figure 1 Trends in published research journal articles on LSA.

in the mainstream IS literature outside Information Retrieval. We anticipate that, as LSA becomes increasingly better-known and implementing software becomes increasingly available, IS researchers will use LSA to analyse large textual databases such as end-user comments or any other form of verbal feedback.

LSA applications relevant to IS research

Kuechler (2007) outlines a number of business and IS applications of analysis of textual data where LSA is applicable. We identify four areas of LSA application that are of particular interest to the IS researchers:

(1) quantitative literature reviews (as done in Landauer et al. 2004, Ord et al. 2005, Larsen et al. 2008, Sidorova et al. 2008, or Hovorka et al. 2009);

(2) analysis of textual data in computer-mediated communication (as done in Abbasi and Chen 2008, or Spomer 2009);

(3) analysis of customer feedback, interviews and free text surveys (as initiated by Coussement and Van den Poel 2008, or Dam and Kaufmann 2008, but more methodological development work needs to be done in that direction);

(4) management of knowledge repositories (as called for in Wei et al. 2008a, Ghose 2009).

The mathematics of LSA

Martin and Berry (2007) provide an introduction to the mathematics of LSA and a small numerical example that illustrates how the analysis works. Valle-Lisboa and Mizraji (2007) provide a rigorous discussion on how LSA detects the underlying topical structure of a document corpus and why LSA's capability of discovering hidden topics allows it to successfully model synonyms, multiple words with similar meaning, and human memory. Similarly, Park and Ramamohanarao (2009) provide a rigorous study on the effect of LSA on term correlation and Larsen and Monarchi (2004) provide an in-depth treatment of the mathematics of LSA and its application in document ("artifact") clustering.

LSA starts with a text quantification method based on what is known as Vector Space Model (VSM) (Salton 1975), where a corpus of d documents using a vocabulary of t terms is used to compile a $t \times d$ matrix **A**, containing the number of times each term appears in each document (term frequencies). Some trivial terms such as "the", "of", etc. (the *stoplist*), are excluded, and some others are consolidated because they share a common stem (term *stemming*, Porter 1980) or some other lexical quality. The frequency counts in **A** typically undergo some transformation (term *weighting*) that penalises common terms and promotes rare ones. After weighting, the term frequencies are typically also normalised so that the sum of squared

transformed frequencies of all term occurrences within each document is equal to one (Salton and Buckley 1988). Subsequently, **A** is subjected to Singular Value Decomposition (SVD),

$$\mathbf{A} = \mathbf{U\Sigma V}^{T}, \quad (1)$$

where **U** are the term eigenvectors, **V** are the document eigenvectors, the superscript ^{T} denotes transposition, **Σ** is a diagonal matrix of singular values (i.e., square roots of common eigenvalues between terms and documents), **UΣ** are the term loadings on the common principal components of terms and documents and **VΣ** are the respective document loadings.

After representing the collection of documents in the space defined by the SVD dimensions, comparisons between documents *i* and *j* can be performed by considering the inner product of rows *i* and *j* of the document loading matrix **VΣ**. Information retrieval queries involve the computation of this kind of similarities between each existing document in the collection and new (query) documents, represented as pseudo-documents, i.e., vectors of term frequencies in the term space, indicating all the terms that appear in each query document (Deerwester et al. 1990). Results of the comparisons among new and existing documents (*query results*) in a matrix format can be obtained by computing the document variance-covariance matrix $\mathbf{R}_{Q} = \mathbf{Q}^{T}\hat{\mathbf{A}}$, where **Q** is the query matrix consisting of pseudo-document vectors, $\hat{\mathbf{A}}$ is the truncated term frequency matrix **A** after SVD and dimensionality reduction and the superscript ^{T} denotes transposition. The inner products that compute the variance-covariance matrix of query results $\mathbf{R}_{Q}$ can be normalized to get the so-called *cosine similarities* between the queries and the documents, which are geometrically interpreted as cosines of angles forming between document and query vectors represented in the space formed by the SVD dimensions.

For purposes of document categorisation, the researcher may want to perform clustering of the documents (Larsen and Monarchi 2004). In traditional multivariate cluster analysis,

observations that are close to each other in space are allowed to form groups (clusters). Similarly, documents that are similar to each other, based on their cosine similarity, form document clusters. Two popular algorithms for document clustering, which both preceded the advent of LSA and have a long history and a broad application domain, are the *K*-means (Witten and Frank 2005 pp 254-260, Johnson and Wichern 2007 pp 696-701), an iterative algorithm that computes cluster centroids, and the expectation maximisation (EM) algorithm (Witten and Frank 2005 pp 265-266), an algorithm for maximum likelihood estimation in much broader contexts and having an older and venerable pedigree.

For purposes of document summarisation, it is desirable to retain a small number of SVD dimensions and represent the documents in the space they define. Following an approach that parallels traditional factor analysis there is the option of rotating term loadings, which allows the researcher to interpret (label) the latent semantic factors. The rotated term loadings are $\mathbf{U\Sigma M}$, where $\mathbf{M}$ is a non-unique rotation matrix having the orthonormality property $\mathbf{MM}^T = \mathbf{I}$, computed by implementing a procedure such as *varimax*. Landauer et al. (2004) claim that "LSA dimensions are fundamentally uninterpretable", however Hu et al. (2007) show how to find a new base with meaningful dimensions and transform the entire LSA space to the new base. In the experience of the authors of the present paper, rotated term loadings obtained through varimax rotations are easily interpretable (see next section for an illustration). The non-uniqueness of the term and document loadings can be easily shown if one observes that $\mathbf{U\Sigma M(U\Sigma M)}^T$, as well as $\mathbf{V\Sigma M(V\Sigma M)}^T$, can reproduce $\mathbf{AA}^T$. Since query results before and after rotations are identical, rotation of the LSA space has no effect on document comparisons.

Table 1 summarises the three LSA methods (cosine similarity, clustering and factor analysis) presented above. For each method, the corresponding analytic goal addressed by the method is also listed on Table 1. Implementing software solutions are presented in the Appendix.

Table 1 Post-LSA methods

Analytic goal	*Post-LSA method*
Document comparisons, document assessment, document classification, coherence among documents	Cosine similarity (queries)
Document categorisation, document summarisation	Clustering, factor analysis

Summary of LSA steps and methodological issues

Figure 2 summarises the individual analytic tasks that are performed as part of LSA. For an informative outline of the various text preparation and analysis steps involved in LSA see also Coussement and Van den Poel (2008). The preparatory steps of term filtering, term stemming and term weighting lay the VSM foundation. The dimensionality reduction step is implemented through an application of SVD. Finally, terms and documents represented in the space defined by the SVD dimensions are analysed by implementing a specific analytic method such as queries (document or term comparisons), clustering, or factor analysis, most of which involve the selection of some type of threshold value. Table 2 summarises five methodological

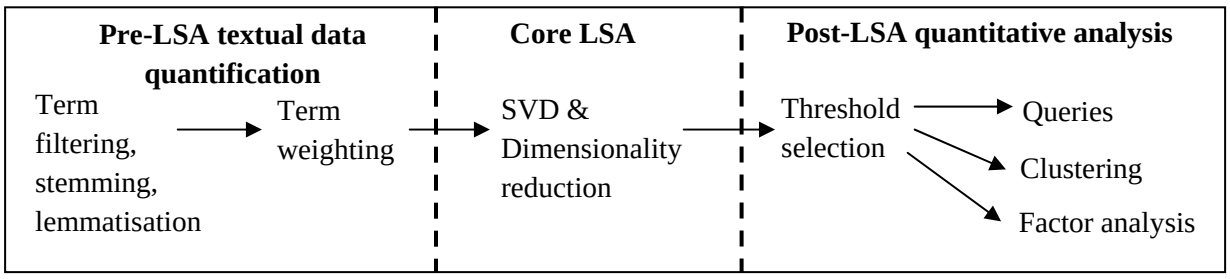


Figure 2 Analytic tasks performed in the context of LSA, including pre-LSA and post-LSA steps

Table 2 Summary of methodological considerations

Consideration	*Methodological choices*
1 Term filtering	Frequency-based *stoplist*
	Stoplist based on explained variance
	Manually selected *stoplist*
	Manually selected *golist*
	Implement Porter Stemmer or Lemmatizer
2 Term weighting	Mostly *TF-IDF* or *log-entropy*
3 Dimensionality reduction	For document comparions: reduce to 100-300 dimensions
	For document summarisation: reduce to 2-200 dimensions
4 Threshold selection	Static or dynamic thresholds for cosine similarity or factor loadings
5 Post-LSA quantitative analysis	Cosine similarities, classification, clustering, factor analysis

considerations related to the aforementioned analytic tasks that constitute the main focus of this paper. For each consideration, corresponding methodological choices are listed. These methodological issues are examined in separate sections of this paper.

Alternative methods

A probabilistic alternative to least-squares-based LSA, called Latent Dirichlet Allocation (LDA) (Blei et al. 2003, Teh et al. 2006, Steyvers and Griffiths 2007), has gained recognition for use in topical analysis. LDA involves estimation of modelling parameters through Markov chain Monte-Carlo (MCMC) simulations. A number of issues related to MCMC implementation (Altman et al. 2004, pp. 118-142) and its ability to address practical applications require further refinement. As a result of its greater mathematical complexity and unresolved issues relevant to application, LDA is beyond the scope of the present work. Another alternative approach that has recently gained popularity in the information retrieval community but is again beyond the scope of this paper is the Nonnegative Matrix Factorisation (NMF) method (Shahnaz et al. 2006, Berry

et al. 2007). In the next section we begin our discussion of the methodological issues in LSA with an elaboration on the choice between clustering and factor analysis.

Clustering versus factor analysis

The extant literature has associated LSA with a number of specific methodological approaches that include information retrieval queries, document classification, or feature extraction. Many quantitative literature overview studies use LSA for document clustering purposes (Landauer et al. 2004, Ord et al. 2005), while some studies (Sidorova et al. 2008) use factor analysis and others (Larsen et al. 2008) use both clustering and factor analysis. In this section we discuss differences and similarities between clustering and factor analysis extensions to LSA.

Document clustering

We start with an illustrative research study, where we quantify abstracts of papers published in the *European Journal of Information Systems* (*EJIS*) using 100 LSA-based dimensions and performing clustering analysis. Since LSA involves a number of methodological issues that require the contribution of a domain expert, we chose the *EJIS* abstracts as our illustrative data in order to give the reader a chance to look at these issues from the domain expert's point-of-view, where the domain is the research published in the *EJIS*. Our own, custom semantic space was created from the corpus of EJIS abstracts. The choice of 100 SVD dimensions is arbitrary, but consistent with common dimensionality choices cited in the literature for document collections of similar size and degree of coherence. The issue of dimensionality selection is addressed in a subsequent section. The 30 clusters resulted from the textual analysis of our data and the examination of a variety of solutions until a unique set of understandable clusters resulted.

Clustering was performed by implementing the Expectation Maximisation (EM) algorithm (Witten and Frank 2005 pp 265-266), a probabilistic version of the K-means clustering algorithm that computes cluster means and standard deviations and assigns terms or documents to clusters by maximising the cluster membership likelihood. Method details are summarised in *Study 1* description below.

> ***Study 1:*** *Abstracts of all research papers published in EJIS in the 1991-2008 period, excluding editorials and commentaries, were collected using an electronic library (EBSCO), for a total of 498 abstracts. An 1873-term by 498-document frequency matrix was composed using a stoplist of 563 common English language terms that carry no specific thematic reference. The TF-IDF term weighting method was applied. Terms were stemmed. Singular Value Decomposition was performed and 100 dimensions were retained. Implementing the Expectation Maximisation (EM) algorithm, 30 clusters were formed. Terms and documents belonging to each cluster were co-examined in order to produce cluster labels.*

The results of Study 1 are summarised in Table 3. The clusters are ranked from largest to smallest. For each cluster, the number of member articles is shown. An effort was made to label the clusters but as the degree of topical coherence varies significantly, the results were not always equally clear. Cluster labelling was done through an iterative process of examining the representative terms and member documents. The authors arrived at the labels as a group, by discussing each label until a consensus was formed. A more appropriate approach would be to involve a number of domain experts (here, IS researchers) and follow the Delphi method for building consensus. Then, in a second round, a number of confederates would match clusters to labels and measures of inter-rater reliability would be obtained (Moore and Benbasat 1991).

Cluster labelling as was done in our study is subject to the possibility of introducing human bias. Some applications such as labelling topics in a discussion group might be best done via an automated approach that would reduce the human bias in representing the groups. For example, Larsen and Monarchi (2004) and Larsen et al. (2008) used Automatic Node Naming for theme label generation. Alternatively, an application such as identifying research areas, or even paradigms, might benefit from an expert researcher's interpretation (Sidorova et al. 2008). Clusters for *EJIS* abstracts labelled with a low degree of confidence are identified in Table 3 by having their label end with a question mark. A major factor that hinders our efforts to label the clusters is the fact that member terms and documents gravitate towards the cluster means, which, in their vector representation, may correspond to linear combinations of more than one semantic space dimensions. This is not necessarily a problem, as the clustering approach aims at an optimal grouping of documents, not at understanding the underlying topical structure.

The largest cluster (C1), consisting of 38 articles, is related to IS development and implementation issues. The second and third clusters are related to knowledge management (C2) and IS and organisational issues (C3), respectively. In general, the clusters reflect research that studies the ways in which information systems get developed and ways in which they interact with organisations, markets and societies, as well as the identity and evolution of the IS discipline. The thematic makeup of EJIS research as presented in our Table 3 should not surprise the reader when contrasted with the 9-topic classification of the entire European IS research presented in Galliers and Whitley (2007), or the 33-topic classification presented in Dwivedi and Kuljis (2008) specifically for *EJIS*. However, direct quantitative comparisons cannot be easily made. For example, Dwivedi and Kuljis (2008) classify 32 articles from the 1997-2007 period as "AI/expert system/neural nets/KM", whereas our Table 3 lists 34 articles from the 1991-2008

period as "Knowledge Management (KM)". These classifications are not necessarily incompatible, but in any case they are not directly equivalent. In another example, Dwivedi and Kuljis (2008) classify 38 articles as "IS research", whereas our Table 3 lists 15 articles as "IS publications", and 10 more as "IS discipline". *IS publications* and the *IS discipline* could be considered as part of *IS research*, but *IS research* might contain additional elements. Further elaboration of the results shown in Table 3 or a formal, quantitative comparison of the thematic make-up presented in Table 3 to other classifications present in the literature, are out of the scope of this paper.

Table 3 Clusters for *EJIS* abstracts

Cluster	*Articles*	*Cluster Label*	*Representative Terms*
C1	38	IS developm. & implementation	system,failure,perspective,development
C2	34	knowledge management	knowledge,integration,task,management,design
C3	33	IS & organisational issues	change,medium,process,requirement,decision
C4	33	outsourcing & applications	outsourcing,technology,application,relationship
C5	31	IT architecture & performance?	firm,network,market,strategy,capability
C6	24	modelling	model,language,domain,represent,conceptual
C7	23	ERP	plan,resource,ERP,enterprise,strategy
C8	23	IT adoption	adoption,influence,indicate,electronic,customer
C9	21	IS methodologies	standard,methodology,provide,stage,compare
C10	17	critical success factors	success,factor,critical,evaluation,literature
C11	17	IT innovation	innovation,mobile,technology,activity,influence
C12	16	social aspects of IS use	team,gender,skill,share,work
C13	16	software products	package,software,developer,determine
C14	15	IS publications	citation,outlet,journal,publication,european
C15	14	IT investment	investment,company,realize,benefit,firm
C16	13	job characteristics	job,user,variable,enterprise,characteristic
C17	12	public sector	e-government,public,government,sector
C18	12	soft systems methodology	soft,SSM,diagram,methodology,debate
C19	11	programming	object,rule,procedure,boundary,program
C20	11	frames and references?	method,reference,frame,actor,scope
C21	11	project management	project,estimate,manager,training,tool
C22	10	IS adoption?	ISD,respond,framework,adopt,explore
C23	10	risk and trust issues	risk,trust,purchase,reduce,consumer
C24	10	IS discipline	disciplinary,discipline,belief,matrix,paradigm
C25	10	online customers	online,customer,intention,service,product
C26	9	idea generation?	imagination,idea,challenge,move,lead
C27	8	GSS	GSS,moderate,leader,play,carry
C28	6	Web	web,site,attribute,model-driven,furthermore
C29	6	social issues?	ICT,telework,meaningful,concept,reflect
C30	4	claims and arguments?	screen,claim,argument,academic,activity

Extraction and labelling of topical factors

The same data used in *Study 1* were now used to produce 30 SVD dimensions. Aiming at an understanding of the latent semantic structure itself, we applied *varimax* rotations on the term loadings, as done in Sidorova et al. (2008). This is a relatively new and emerging technique that provides some future research opportunities. Varimax rotations are common in traditional factor analysis. They simplify the ability to understand the factor loadings, by making as many of them as possible large in size, and as many as possible small in size. After varimax rotations there is a clearer association between factors and loading variables and that makes factor labelling more straightforward. In our case, the rotated term loadings produce factors that are naturally easy to interpret given that the factor space and the researcher "speak the same language," i.e., the terms are a set of variables for the factor space and a language vocabulary for the researcher. In order to maintain the documents' relationship with the factor space, documents were also rotated by applying the same rotation matrix used to rotate the terms. Regarding information retrieval queries, the rotated factor space maintains its ability to produce identical similarity metrics (cosine similarities) for terms and documents. Method details are summarised in *Study 2* description below.

> ***Study 2:*** *Using the same data as in Study 1 (498 EJIS abstracts) and the same 1873-term by 498-document frequency matrix, Singular Value Decomposition was performed and 30 dimensions were retained. After performing varimax rotations on the term loadings matrix and replicating the same rotations on the document loadings, lists of high-loading terms and high-loading documents were compiled. High-loading terms and documents were co-examined in order to produce factor labels.*

The results of Study 2 are summarised in Table 4, where factors are rank-ordered based on the corresponding principal components. In a manner similar to what is done in traditional factor analysis, the quality of the solution can be judged according to the proverbial "WOW criterion" (Johnson and Wichern 2007, p. 526): a good solution will always make good sense to the researcher. In our case the 30-factor solution produces factors with good correspondence to the 33-topic classification in Palvia et al. (2004), which was used to classify *EJIS* articles by Dwivedi and Kuljis (2008). An elaborate discussion of the IS research topics listed in Table 4 is peripheral to the focus of this paper, but one can recognise well-established research topics such as *soft systems methodology* (F30.1), *enterprise resource planning* (F30.7), *electronic data interchange* (F30.8), *outsourcing* (F30.12) and *agile software development* (F30.21). The factors were labelled easily and without controversy following a procedure similar to the one described in the clustering analysis section. Comparing the labels produced by clustering and factor analysis (Studies 1 and 2), we observe a significant amount of conceptual overlapping. For example, the clustering approach produced a *soft systems methodology* cluster (C18 in Table 3) that contained 12 articles, while the factor analysis approach produced a *soft systems methodology* factor (F30.1 in Table 4) with 13 high-loading articles that included all 12 from cluster C18. Cluster C18 is represented by the highest-likelihood member terms *soft*, *SSM*, *diagram*, *methodology*, and *debate*. Factor F30.1 is described by the top-loading stemmed terms *SSM*, *soft*, *methodologi-*, *rich*, *picture-*, and *system*. The two sets of terms imply a conceptual makeup that is similar, even though not identical, in both cases corresponding to the *soft systems methodology* research topic. Tables 3 and 4 also present a number of differences. Some of them may be due to the difference in number of underlying SVD dimensions, 100 in Study 1, versus 30 in study 2. Indeed, when we tried generating 30 clusters based on only 30 underlying SVD

dimensions, we observed that the clustering results were quite different from the clustering results obtained by retaining 100 dimensions, shown on Table 3. Furthermore, the clusters based on 30 dimensions had a larger proximity to the 30 factors shown on Table 4 than the clusters based on 100 dimensions had to the same 30 factors. The reason might be related to the complexity introduced by the 100 dimensions to the nature of the clusters shown on Table 3. While it is clear that more research is needed on the subject of comparing clustering to factor

Table 4 Topical factors for *EJIS* abstracts

Factor	*Articles*	*Factor Label*	*High-Loading Terms*
F30.1	13	soft systems methodology	SSM,soft,methodologi,rich,pictur,system
F30.2	28	firm performance	firm,perform,sophist,align,strateg
F30.3	20	IS publications	journal,European,quality,publish,citat,rank
F30.4	25	social perspectives	materi,action,social,human,view,theori
F30.5	22	innovation	Innov,diffusion,perceive,adop,barrier
F30.6	21	software	softwar,product,packag
F30.7	15	ERP	ERP,enterpr,implement,expect,involv
F30.8	23	EDI	inter,network,EDI,coordin,integr,market
F30.9	22	domain and data models	method,domain,design,relationship,data
F30.10	20	IS evaluation	benefit,realiz,invest,compani,expect,success
F30.11	27	Ebusiness	firm,ebusi,custom,competit,SME
F30.12	12	outsourcing	outsource,vendor,contract,sector,IT
F30.13	12	mobile services	mobil,service,user,switch,tempor
F30.14	18	public sector	govern,public,sector,
F30.15	15	instrument development	valid,measur,construct,dimen,perform
F30.16	17	project failure/success	failure,project,success,ISD,implement
F30.17	17	online consumer	trust,consum,onlin,risk,ecommerc
F30.18	17	IS planning	plan,strategi,strateg,implement
F30.19	12	team projects	team,project,locat,global,collabor,distribut
F30.20	6	DSS	DSS,quality,assess,decision
F30.21	13	agile software development	agil,ISD,chang,method
F30.22	10	security	securi,risk,conceptu,issue,control
F30.23	19	IT and groups	decision,role,influenc,task,GSS
F30.24	14	critical success factors	success,critic,factor,taxonomi,CSF
F30.25	18	IS investment evaluation	evalu,invest,method,assess
F30.26	8	IT jobs	job,comput,USA,personnel
F30.27	15	imagination and knowledge	knowledge,imagin,SME,expert,Weick
F30.28	9	IS discipline	discipline,claim,enterpr,polit,academ
F30.29	11	knowledge management	knowledge,learn,AR,academ,KM
F30.30	19	Web site development	Web,custom,risk,content,internet,implement

analysis results, our observation is quite interesting, given that most clustering efforts are based on spaces that retain a much higher dimensionality than the number of clusters they aim to create. As we compare the results of the two approaches (Tables 3 and 4), it is important to realise that such a comparison is potentially biased in favour of the factor analysis approach: cognitive science theory has proposed that our human brain is wired to understand the latent semantic dimensions because it operates by them (Landauer 2007). LSA identifies a common set of patterns in textual data and presents them to the researcher in the form of high-loading terms and documents that make the patterns more easily discernable than from the raw textual data. Conceptual knowledge is gained when LSA is used to synthesise the text into structured factors and a relevant meaning is discerned from the textual data and assigned to those factors.

Comparing the clustering and factor analysis approaches

Figure 3 illustrates similarities and differences in clustering (Figure 3a) and factor analysis (Figure 3b) results, based on a hypothetical situation where eight documents (*d1-d8*) are

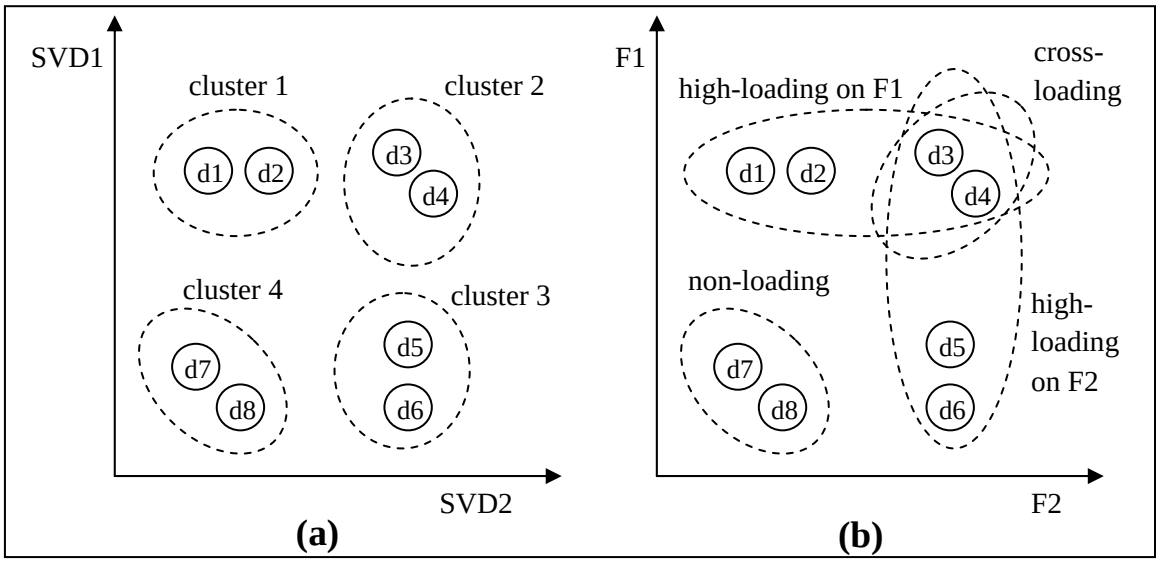


Figure 3 Document grouping in clustering (a) and factor analysis (b) approaches.

represented in a latent semantic space of two dimensions. The clustering approach produces four clusters, as shown in Figure 3a. Each document is forced to join exactly one cluster. This was also the case with our Study 1, where all 498 abstracts were required to participate in exactly one cluster (the article counts in Table 3 add up to 498.) The factor analysis approach produces a set of high-loading documents for factor F1, {*d1*, *d2*, *d3*, *d4*}, and a set of high-loading documents for factor F2, {*d3*, *d4*, *d5*, *d6*}. Our Study 2 also produced a number of cross-loading documents: the article counts in Table 4 still add up to 498, but 99 articles loaded on more than one factor, 25 of which on more than 2 factors. In other words, some of these 99 articles were double-counted, some triple-counted, etc., leaving a corresponding number of 129 articles failing to load on any of the 30 factors. Back to Figure 3, moving from a clustering approach to a factor analysis approach does not necessarily result in a loss of information because set {*d3*, *d4*} is detected based on factor cross-loading status and set {*d7*, *d8*} is detected based on non-loading status. Yet, the factor analysis approach focuses on the topical structure itself, i.e., the extraction and understanding of factors F1 and F2, while the clustering approach focuses on the identification of groups of similar documents, i.e., clusters 1-4. Researchers should consider their research questions and then decide which approach might serve them best. As our Study 1 and Study 2 demonstrated, clustering and factor analysis results are not likely to be identical. We conclude our comparison between the clustering and factor analysis approaches to LSA, by noting that, if the purpose of a literature summarisation study is the identification of research groups, or communities, document clustering as done by Larsen et al (2008) is probably more appropriate. If, however, the purpose is the identification of research dimensions or themes that underlie and transcend the research documents, factor labelling as done by Sidorova et al. (2008) is probably

more appropriate. We summarise our recommendation regarding methodological extensions to LSA in general, as follows:

> ***Recommendation 1:*** *Researchers should select among classification, clustering and factor analysis extensions to LSA, whichever is more appropriate for addressing their research questions:*
>
> - *if the research goal is to match documents to pre-existing categories they should perform document classification extensions to LSA;*
> - *if the goal is to generate new, data-driven document groups they should perform document clustering;*
> - *if the goal is to understand the latent structure of their corpus, they should perform factor analysis extensions to LSA.*

In the rest of the paper we will focus on the factor analysis approach. This choice was not made because we wish to underplay clustering or document comparison approaches, but rather because, as we continue illustrating certain methodological details that are common to all these approaches, we find the factor analysis approach easier for our readers to follow.

A note on threshold selection

Table 4 lists the number of relevant documents (*EJIS* articles) for each factor, i.e., the number of documents that load sufficiently high on each topical factor and, at the same time, maintain some recognisable topical proximity with the factor. How is such proximity determined? One option is to use 0.40 or some other loading threshold that is commonly used in traditional Factor Analysis. We caution against such generalisations. In our Study 2, for our illustration purposes we used a heuristic: each article should relate on average, to one topic, therefore, each document should

load, on average, on one factor. This resulted in a document loading threshold of 0.179. An examination of the weakest loadings for each factor revealed that documents with a loading of 0.18 maintained, in the vast majority of the cases, a reasonably strong relationship to their respective topical factor. Even though our heuristic produced acceptable results, we recommend that, until a well-established method for loading threshold selection is in place, researchers manually select a threshold for each factor separately:

> ***Recommendation 2:*** *Researchers taking the factor analysis approach to LSA should not apply 0.40 or some similarly preset loading threshold, but instead apply an empirically derived threshold, validated by a domain expert because thresholds as low as 0.18 were found acceptable.*

This recommendation is in line with the recommendation in Penumatsa et al. (2006) not to use 0.65 as a fixed threshold for cosine similarity measures in document comparisons but, instead, to allow for dynamic adjustment of the threshold, since cosines as low as 0.40 can produce good results, in terms of information retrieval *recall* and *precision* values. It is also in line with research findings indicating that cosine similarity increases monotonically with document size (Hu et al. 2007). The issues of loading threshold selection, as well as cosine threshold selection, underline the need to have a domain expert qualitatively validate the LSA results.

Term filtering

Term selection remains an open issue. The need to reduce the term dimensionality arises because of the desire to achieve computational efficiency, as well as to avoid overfitting of the semantic space. To serve such legitimate needs, a common approach is to filter terms that appear only a few times in the entire collection of documents (*frequency filtering*). A variety of term frequency

thresholds have been employed in the literature. Griffiths and Steyvers (2004) retain terms that appear in at least five documents. Ord et al. (2005) reduce their 24,850 initial terms to just 739 by keeping terms that appeared at least 100 times in their collection of documents. Sidorova et al. (2008) retain terms that have a high capacity to explain variability (*communality filtering*). We encourage the researcher to explore these choices, but we caution against undisclosed term selection practices, as there is strong potential for manipulation. Term selection can also involve difficulties when dealing with data sets containing proper names, numbers, abbreviations, SMS-speak, slang, acronyms etc. as well as with very small text units (twitters, single sentences). In addition, term stemming can sometimes combine words in a non-meaningful way. For example, Porter stemmer (Porter 1980) combines *community* and *communication* into the single stem "*commun-*". Alternative approaches such as lemmatisation, which conflates words after identifying their corresponding part of speech, are proposed to better discriminate between words that have different meanings and are sometimes cited as superior to stemming (Lifchitz et al. 2009). For example, it may be desirable to distinguish between the verb *address* as in "addressing the issue" versus the noun *address* as in "E-mail addresses". Some of these issues can be addressed via careful preprocessing at the pre-LSA steps. The reader is encouraged to visit the morphadorner.northwestern.edu site for a collection of related text processing utilities.

In order to illustrate the importance of vocabulary selection and its impact on the final LSA results, we now adopt an extreme approach where terms are selected manually and a special-purpose vocabulary is compiled. Analysing again the set of 498 *EJIS* abstracts used in our previous studies, we force LSA to ignore terms related to topics and only consider terms related to research methods. Being aware of the research methodology classification used in Palvia et al. (2004) and Dwidevi and Kuljis (2008), we first compiled a list of 230 methodology-

related terms, such as *action*, *analysis*, *case*, *data*, *hermeneutics*, *instrument*, *interpretive*, *study*, *survey*, *theory*, etc. We then used these 230 terms as the "*golist*" (also called *startlist*), i.e., we forced our Vector Space Model to use these particular 230 terms to describe the 498 documents and produced a 12-factor solution. Method details are summarised in *Study 3* description below.

> ***Study 3:*** *Using the same data as in Study 1 (498 EJIS abstracts), a 230-term by 498-document frequency matrix was composed using a golist of 230 methodology-related terms. TF-IDF term weighting was applied. Singular Value Decomposition was performed and 12 dimensions were retained. After performing varimax rotations, factor labelling was attempted based on high-loading terms and documents.*

The methodological choices made in study 3 relevant to the specific 230 terms used as a *golist* and the dimensionality choice were the result of several iterations. Labels for the 12 factors are presented in Table 5. The solution provided, while not necessarily final, is a good start. The extracted factors compare well to the 16-method classification used for EJIS in Dwidevi and Kuljis (2008), with 9 out of 12 factors (75%) overlapping. Factors F12.8: *methodology* and F12.12: *quantitative analysis* are probably too generic to be useful, therefore a successive iteration could try to dissolve them by eliminating their high-loading terms from the effective vocabulary (golist), split them by increasing the dimensionality to 13 factors, or merge them by reducing the dimensionality to 11 factors. Another possibility is to apply log-entropy term weighting (see next section). Since the crafting of the definitive list of methodological factors in EJIS abstracts is not the intent of the present paper, we do not go any farther in this direction. Our main purpose was to provide an extreme example that illustrates both the research opportunity and the potential for manipulation when it comes to term selection. Moreover, such

potentials are further leveraged by interactions between term selection, term weighting and dimensionality choices. These interactions will be revisited in subsequent sections.

Table 5 Methods view of the *EJIS* abstracts: 12-factors, TF-IDF weighting

Factor	*Articles*	*Methods-factor label*	*Selected high-loading terms*
F12.1	42	theoretical model/framework	framework,theori,conceptu,model
F12.2	52	case study	casestudi
F12.3	52	theory/opinion	social,argu,theori
F12.4	29	action research	action
F12.5	48	literature review	literatur,review
F12.6	34	data collection	data,collect,model,compar
F12.7	45	survey	survei
F12.8	35	methodology	studi,argu,mean,sampl,compar
F12.9	36	instrument development	measur,valid,construct,instrum
F12.10	46	secondary data analysis	analyz,theori,compar,solution
F12.11	39	qualitative research	argu,debat,code,analyz,content
F12.12	40	quantitative analysis	analysi,econom,solution,model

The main purpose of Study 3 was to illustrate how LSA produces a dramatically different set of dimensions when alternative term frequency matrices are compiled from the same corpus by applying alternative lists of terms. This should not come as a surprise, as conceptual meaning in human communication is transcribed through words: emphasising alternative sets of words would result in emphasising alternative sets of concepts. Our recommendation regarding term selection is summarised below.

> ***Recommendation 3:*** *Researchers should disclose the terms used (golist) or the terms filtered out (stoplist) because the vocabulary of terms used in LSA is critical in determining the analysis results: concepts can be added or removed from the latent semantic space by including or excluding terms related to those concepts.*

Term weighting

The problem of finding the optimal weighting method for transforming the term frequencies is addressed extensively in the information retrieval literature. Two of the most widely used

transformations are the Inverse Document Frequency transformation, commonly referred to as *TF-IDF*, (Salton 1975, Husbands et al. 2001, Han and Kamber 2006 p.619) and the *Log-Entropy* transformation (Dumais 1991, Chew et al. 2007), which was found to outperform TF-IDF for purposes of information retrieval and document classification. For purposes of document summarisation, however, where the issue of optimal term weighting largely remains exploratory, it is worthwhile to try more than one transformation to ensure interpretative consistency. In the IS literature, TF-IDF weighting has been used in Coussement and Van Den Poel (2008), Larsen et al. (2008), Sidorova et al. (2008) and Wei et al. (2008a, 2008b). Log-Entropy weighting utilisation in the IS literature has been scarce.

Revisiting our illustration study 2, we re-analysed the 498 EJIS abstracts keeping all 1873 terms after applying log-entropy weighting and we extracted 30 factors. Out of the 30 topics distilled using TF-IDF weighting, most topics were retained under log-entropy weighting. The four topics that are identified only under log-entropy weighting are *frameworks*, *implementation*, *prototype design* and *communication networks*. Topics such as *IS publications* and *social perspectives* that were identified using TF-IDF, were dissolved when log-entropy was used. In order to further investigate the effect of term weighting on factor formation, correlations among TF-IDF term communalities, log-entropy communalities, and term frequencies were calculated. The results show that term frequencies are more correlated to log-entropy communalities (Pearson's correlation coefficient $r = 0.66$), than to TF-IDF communalities ($r = 0.43$).

Revisiting our illustration study 3, we re-analyzed the same 498 *EJIS* abstracts after applying log-entropy weighting. Once again, we extracted 12 factors. The factor labels shown in Table 6 were once again produced after co-examination of high loading terms and documents (abstracts) even though, due to space limitations, Table 6 shows only the top loading terms. One

noticeable difference between the factors in Tables 5 and 6 is the splitting of the theoretical model/framework factor into separate factors for *model, framework,* and *theoretical model*. The new factors are built around fewer terms. They typically load on one term with a loading value around 2.0 or more, followed by weaker loadings around 0.6 or less.

Our findings reveal a potential weakness of the log-entropy transformation: factors are potentially biased towards high-frequency key terms. Interestingly, when analysing the EJIS article titles (as opposed to abstracts), we found that the log-entropy transformation worked better. This finding suggests that log-entropy works better at obtaining a small number of factors that rely on a few, frequently used terms. Article titles or short text messages may yield better results when a log-entropy transformation is used, because they stay closer to the "periphery" of the language structure, i.e., to a few relatively frequent words whose presence versus absence makes a critical difference. In contrast, TF-IDF appears to be better at discovering patterns in the "core" of the language, i.e., it identifies larger groups of terms which tend to appear all together in moderate frequencies. For example, while both weighting methods resulted in the extraction of a *case study* (see F12.2 in Table 5 and F12.3 in Table 6) and an *action research* factor (see F12.4 in Table 5 and F12.7 in Table 6), only the TF-IDF weighting method resulted in the extraction of a combined *theoretical model/framework* factor (F12.1 in Table 5).

As a general remark, we would like to point out that a direct examination and labelling of the concepts that correspond to the dimensions of the latent semantic space in the way that is presented here is not very common in the LSA literature. Therefore, researchers may often not be fully aware of the semantic space configuration changes that result from term weighting choices such as the ones presented in this section, or term selection choices such as those presented in the previous section. We conclude with the following recommendation:

Table 6 Methods view of the *EJIS* abstracts: 12 factors, Log-Entropy weighting

Factor	*Articles*	*Methods-factor label*	*Selected high-loading terms*
F12.1	70	model	model,data,conceptu,semant,framework
F12.2	56	framework	framework,conceptu,literatur,theoret
F12.3	48	case study	casestudi
F12.4	44	data collection	data,compar,collect,rate,sampl
F12.5	36	theoretical model	theori,model,survei,theoret
F12.6	39	instrument development	measur,valid,construct,model,instrum
F12.7	31	action research	action
F12.8	41	theory, opinion, literature review	social,review,argu,literatur,theori
F12.9	37	analysis	analysi,social,econom,studi,
F12.10	38	survey, literature review	literatur,survei,variabl
F12.11	32	qualitative & conceptual research	argu,debat,solution,studi,conceptu
F12.12	27	analysis, content analysis	analyz,content,review

Recommendation 4: *Researchers are encouraged to consider alternative transformations, such as TF-IDF or Log-Entropy, and select the method that is most closely aligned with the research question and intent. Based on our experiments:*

- *TF-IDF was more appropriate when the intent was to represent documents in a relatively conceptual and complex semantic space;*
- *Log-Entropy was more appropriate when the intent was to represent documents in a semantic space built around a few key terms.*

Dimensionality reduction

The problem of selecting an appropriate number of latent semantic dimensions was dealt with empirically and remains open. Bradford (2008) summarises the optimal factor numbers used in 49 published LSA studies (see Table 1 in Bradford 2008), ranging from 6 to over 1,000. For collections of about 5,000 terms by 1,000 documents, a choice of about 70 to 100 dimensions is frequently cited as optimal (Deerwester et al., 1990). Efron (2005) selects the number of

dimensions based on non-parametric confidence intervals obtained through simulations and bootstrapping. Interestingly, for collections of similar size, his method selects values in the range of 80 to 100. On the other hand, Doxas et al. (2010) show with the help of simulations that as they construct prose, authors traverse a semantic space of approximately eight dimensions regardless of language or genre, implying that latent semantic dimensionality for most corpora may be surprisingly low. Zhu and Ghodsi (2006) propose a much simpler approach to dimensionality selection which applies a log-likelihood test on the eigenvalues, seeking an "elbow" point on the scree plot, called Profile Likelihood Test (PLT). Traditional factor analysis approaches such as 85% of total variance explained or the Kaiser-Guttman rule of selecting components whose eigenvalues are greater than the mean eigenvalue, typically select a larger number of components. We caution against such blanket approaches because the correct choice depends on the specific corpus and also the analytic goal. Researchers who pursue a simplification of the original, redundant textual space for purposes of calculation efficiency and still wish to explain a large percentage of variance may want to select a higher dimensionality than researchers who try to distil the semantic core of such space at an abstract, high level that explains only a small fraction of total variance in the textual data.

Our discussion of dimensionality selection continues with the introduction of an illustration study, summarised below.

> ***Study 4:*** *Abstracts of three special issues articles of EJIS published in issues 15(1), 15(2), and 16(6) (22 abstracts) were collected. The 261-term by 22-document frequency matrix was weighted using a TF-IDF transformation. The Profile Likelihood Test performed on the eigenvalues indicated 7 as the optimal dimensionality. The factor analysis approach to LSA was implemented and two*

solutions were obtained. In one solution, seven SVD dimensions were retained, rotated and labelled as in Study 2. In the second solution, only three SVD dimensions were retained and rotated.

Several studies (for example, Haley et al. 2007, Bradford 2008) show that dimensionality selection remains an unresolved methodology issue. For the purpose of illustration we provide details on one of the approaches suggested in the literature. We illustrate how to use a scree plot to assist in the selection of the appropriate number of dimensions. Figure 4 shows the scree plot for the obtained 22 principal components. Our implementation of the PLT yielded a highly significant (p-value = 0.0038) point estimate at principal component k = 7, where the maximum log-likelihood (Q_n = 33.04) was attained (see Zhu and Ghodsi 2006 for details). Table 7 lists the high-loading articles of each of the seven corresponding LSA factors, together with the absolute values of their factor loadings.

Table 8 lists the high-loading articles for the corresponding three-factor solution. Based on high-loading terms, the three factors appear to be related to *business agility*, *healthcare IS*, and *action-oriented research*, respectively. The high-loading documents as presented in Table 8, confirm this contention. At first look, this 3-factor solution appears more appealing, since it

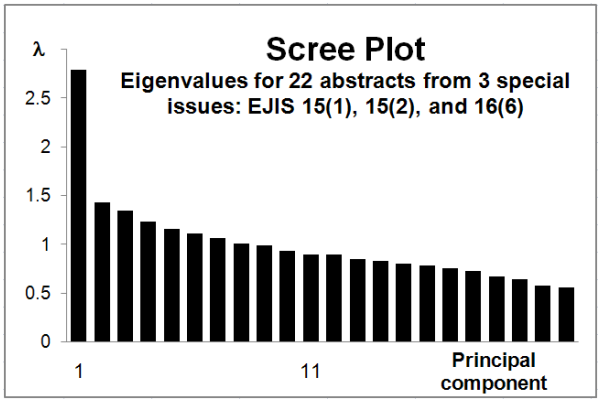


Figure 4 Scree plot for the 22 abstracts analysed in study 4.

Table 7 Seven topical factors for 22 special-issue *EJIS* abstracts

Author	Special Issue	EJIS Ref.	Factor Loadings						
			F7.1	F7.2	F7.3	F7.4	F7.5	F7.6	F7.7
Overby et al.	Business Agility	15(2), 120-131	0.706						
Hovorka & Larsen	Business Agility	15(2), 159-168	0.585						
van Oosterhout et al.	Business Agility	15(2), 132-145	0.542						
Lyytinen & Rose	Business Agility	15(2), 183-199	0.426						
Rittgen	Action-Oriented Res.	15(1), 70-81		0.700					
Ågerfalk et al.	Action-Oriented Res.	15(1), 4-8		0.565					
Gasson	Action-Oriented Res.	15(1), 26-41		0.477					
Andersen	Action-Oriented Res.	15(1), 9-25		0.391					
Karlsson & Wistrand	Action-Oriented Res.	15(1), 82-90			0.783				
Yetim	Action-Oriented Res.	15(1), 54-69			0.617				
Bondarouk	Action-Oriented Res.	15(1), 42-53				0.636			
Reardon & Davidson	Healthcare IS Res.	16(6), 681-694				0.605			
Jensen & Aanestad	Healthcare IS Res.	16(6), 672-680				0.528			
VanAkkeren&Rowlands	Healthcare IS Res.	16(6), 695-711				0.507			
Börjesson et al.	Business Agility	15(2), 169-182					0.594		
Fitzgerald et al.	Business Agility	15(2), 200-213					0.588		
Holmqvist & Pessi	Business Agility	15(2), 146-158					0.470		
van Oosterhout et al.	Business Agility	15(2), 132-145					0.383		
Cho & Mathiassen	Healthcare IS Res.	16(6), 738-750						0.727	
Lee & Shim	Healthcare IS Res.	16(6), 712-724						0.671	
Krogstie et al.	Action-Oriented Res.	15(1), 91-102						0.352	
Bhattacherjee & Hikmet	Healthcare IS Res.	16(6), 725-737							0.731
Klein	Healthcare IS Res.	16(6), 751-760							0.672

produces cleaner factors. A more careful examination, however, will reveal that the 3-factor solution left out the article authored by Tanya Bondarouk (2006). This omission is offset by a double-count: Hovorka and Larsen (2006) cross-load on both the *business agility* and the *healthcare IS* factors. This is probably attributed to the fact that Hovorka and Larsen (2006) study agile adoption practices using a language that talks about "*adoption* of information technology (IT)-based *innovations*", in a way that gets associated with the discussion of "telehealth *innovation*" in Cho and Mathiassen (2007) and "*adoption* of healthcare information systems" in Blegind Jensen and Aanestad (2007). Comparing the two solutions presented in

Tables 7 and 8, we observe that the 3-factor solution, based strictly on the top three underlying SVD dimensions is better at summarising the underlying three special topics in the form of extracted latent factors, at the expense of not representing all the articles. We suggest that the researcher examine solutions based on his or her expert knowledge of the underlying theory because our comparison illustrates the trade-off between fitting the theory versus explaining all the variance. Because the researcher cannot directly observe the LSA dimensions, in order to understand the latent semantic space, the researcher has to rely on post-LSA interpretation. Clustering and factor analysis are aids in doing so. As a result, these methodologies are important in providing insight into the LSA dimensions. We summarise our recommendation regarding dimensionality selection below.

Table 8 Three topical factors for 22 special-issue *EJIS* abstracts

Author	Special Issue	EJIS Reference	Factor Loadings		
			F3.1	F3.2	F3.3
van Oosterhout et al.	Business Agility	15(2), 132-145	0.654		
Fitzgerald et al.	Business Agility	15(2), 200-213	0.583		
Overby et al.	Business Agility	15(2), 120-131	0.583		
Holmqvist & Pessi	Business Agility	15(2), 146-158	0.463		
Lyytinen & Rose	Business Agility	15(2), 183-199	0.447		
Börjesson et al.	Business Agility	15(2), 169-182	0.443		
Hovorka & Larsen	Business Agility	15(2), 159-168	0.427		
Reardon & Davidson	Healthcare IS Res.	16(6), 681-694		0.663	
Lee & Shim	Healthcare IS Res.	16(6), 712-724		0.504	
Van Akkeren & Rowlands	Healthcare IS Res.	16(6), 695-711		0.479	
Cho & Mathiassen	Healthcare IS Res.	16(6), 738-750		0.413	
Klein	Healthcare IS Res.	16(6), 751-760		0.388	
Jensen & Aanestad	Healthcare IS Res.	16(6), 672-680		0.329	
Hovorka & Larsen	Business Agility	15(2), 159-168		0.314	
Bhattacherjee & Hikmet	Healthcare IS Res.	16(6), 725-737		0.306	
Karlsson & Wistrand	Action-Oriented Res.	15(1), 82-90			0.581
Ågerfalk et al.	Action-Oriented Res.	15(1), 4-8			0.514
Yetim	Action-Oriented Res.	15(1), 54-69			0.498
Gasson	Action-Oriented Res.	15(1), 26-41			0.460
Rittgen	Action-Oriented Res.	15(1), 70-81			0.455
Krogstie et al.	Action-Oriented Res.	15(1), 91-102			0.331
Andersen	Action-Oriented Res.	15(1), 9-25			0.253

> ***Recommendation 5:*** *Researchers are encouraged to explore and report alternative dimensionalities and perform sensitivity analysis as well as qualitative assessments that link the results to underlying theory, because the appropriate number of SVD dimensions, clusters, factors, or predefined categories remains an open issue.*

Conclusion

In this paper we discussed various methodological issues that arise in the context of Latent Semantic Analysis, an emerging quantitative method for the analysis of textual data. Our main recommendations are summarised in Table 9.

In conclusion, we believe that while LSA is very broadly applicable, it has numerous applications that are of potential interest to IS researchers that have not yet materialised because of lack of familiarity with the methodology. Such applications include the analysis of leadership vision statements, corporate announcements, regulatory body statements, expert assessment notes, customer feedback comments, open-ended surveys, text messages, Web content, news stories, and IS publications. Thus, the application domain for LSA includes textual data generated in individual, organisational, and societal contexts of developing, using, and studying Information Systems. As a final remark, we would like to emphasise the importance of intelligent interpretation of the results of the quantitative analysis on the part of the researcher. LSA is a quantitative technique and, as such, requires some intelligent selection of important parameters on the part of the researcher. However, a solution that has been fine-tuned by addressing effectively the methodological issues discussed in this paper will still need to make good sense to the researcher, and this is where quantitative analysis and subjective judgement

Table 9 Summary of five methodological recommendations

Issue		*Recommendation*
LSA extension: a number of post-LSA quantitative analysis methods have been used in the literature, including document comparisons, clustering, classification, categorisation, and factor analysis	R1	Select among classification, clustering and factor analysis extensions to LSA, whichever is more appropriate for addressing your research questions: • if the research goal is to match documents to pre-existing categories, you should perform document classification extensions to LSA; • if the goal is to generate new, data-driven document groups, you should perform document clustering; • if the goal is to understand the latent structure of their corpus, you should perform factor analysis extensions to LSA
Cosine and loading thresholds: a variety of thresholds have been used in the literature, some of them as low as 0.18	R2	Do not apply preset loading thresholds such as 0.40, but instead apply an empirically derived threshold, validated by a domain expert
Term selection: the vocabulary of terms used in LSA can be critical in determining the analysis results	R3	Disclose the terms used (*golist*) or the terms filtered out (*stoplist*)
Term weighting: no term-weighting method is known to be universally best	R4	Consider alternative transformations, such as TF-IDF or Log-Entropy, and select the method that is most closely aligned with the research question and intent. Based on our experiments: • TF-IDF was more appropriate when the intent was to represent documents in a relatively conceptual and complex semantic space; • Log-Entropy was more appropriate when the intent was to represent documents in a semantic space built around a few key terms
Dimensionality selection: the estimation of an appropriate number of SVD dimensions, clusters, factors, or predefined categories remains an open issue	R5	Explore and report alternative dimensionalities and perform sensitivity analysis as well as qualitative assessments that link the results to underlying theory

intercept. Conversely we believe that, while content analytic approaches will largely retain their traditional qualitative nature for a while, their methodological mix will soon acquire a strong quantitative component involving methods such as LSA or LDA. We have only seen the introduction of such methodological approaches in IS research and the application and development remains a fertile area for future work. The authors hope that the present paper will encourage research in these methodologies.

References

ABASI A and CHEN H (2008) CyberGate: A design Framework and System for Text Analysis of Computer-Mediated Communication. *MIS Quarterly* 32(4), 811-837.

ALTMAN M, GILL J and MCDONALD M (2004) *Numerical Issues in Statistical Computing for the Social Scientist*. Wiley Series in Probability and Statistics.

BAJWA IS, SAMAD A AND MUMTAZ S (2009) Object Oriented Software Modeling Using NLP Based Knowledge Extraction. *European Journal of Scientific Research* 35(1), 22-33.

BARRET MI (1999) Challenges of EDI adoption for electronic trading in the London Insurance Market. *European Journal of Information Systems* 8(1), 1-15.

BERRY M, DUMAIS S and O'BRIEN G (1995) Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review*, 37(4), 573-595.

BERRY MW, BROWNE M, LANGVILLE AN, PAUCA VP and PLEMMONS RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis* 52(1), 155-173.

BLEGIND JENSEN T and AANESTAD M (2007) Hospitality and hostility in hospitals: a case study of an EPR adoption among surgeons. *European Journal of Information Systems* 16(6), 672-680.

BLEI DM, NG AY and JORDAN MI (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

BONDAROUK TV (2006) Action-oriented group learning in the implementation of information technologies: results from three case studies. *European Journal of Information Systems* 15(1), 42-53.

BRADFORD RB (2008) An empirical study of required dimensionality for large-scale latent semantic indexing applications. *CIKM '08: Proceeding of the 17th ACM conference on information and knowledge management*, ACM, New York, 153-162.

CHEW P, BADER B, KOLDA T and ABDELALI A (2007) Cross-Language Information Retrieval Using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD* (GAFFNEY S, Ed), pp 143-152, ACM Publications, Baltimore, Maryland.

CHO AND MATHIASSEN (2007) The role of industry infrastructure in telehealth innovations: a multi-level analysis of a telestroke program. *European Journal of Information Systems* 16(6), 738-750.

COUGER JD and O'CALLAGHAN R (1994) Comparing the motivations of Spanish and Finnish computer personnel with those of the United States. *European Journal of Information Systems* 3(4), 285-291.

COUSSEMENT K and VAN DEN POEL D (2008) Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems* 44(4), 870-882.

DAM G and KAUFMANN S (2008) Computer assessment of interview data using latent semantic analysis. *Behavior Research Methods* 40(1), 8-20.

DAMSGAARD J and TRUEX D (2000) Binary trading relations and the limits of EDI standards: the Procrustean bed of standards. *European Journal of Information Systems* 9(3), 173-188.

DEERWESTER S, DUMAIS S, FURNAS G, LANDAUER T and HARSHMAN R (1990) Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41(6), 391-407.

DOXAS I, DENNIS S, OLIVER WL (2010) The dimensionality of discourse. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 107, 4866-4871.

DUMAIS ST (1991) Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers* 23(2), 229-236.

DUMAIS ST (2004) Latent Semantic Analysis. *Annual Review of Information Science and Technology* 38, 189-230.

DUMAIS ST (2007) LSA and Information Retrieval: Getting Back to Basics. In *Handbook of Latent Semantic Analysis* (LANDAUER TK, MCNAMARA DS, DENNIS S and KINTSCH W, Eds), pp 293-322, Lawrence Erlbaum Associates, Mahwah, New Jersey.

DWIVEDI YK and KULJIS J (2008) Profile of IS research published in the European Journal of Information Systems. *European Journal of Information Systems* 17(6), 678-693.

EFRON M (2005) Eigenvalue-Based Model Selection During Latent Semantic Indexing. *Journal of the American Society for Information Science and Technology*, 56(9), 969-988.

FRANZOSI R (2004) *From Words to Numbers: Narrative, Data, and Social Science*. Cambridge University Press, Cambridge, United Kingdom.

GALLIERS RD and WHITLEY EA (2007) Vive les differences? Developing a profile of European information systems research as a basis for international comparisons. *European Journal of Information Systems* 16(1), 20-35.

GHOSE A (2009) Internet Exchanges For Used Goods: An Empirical Analysis Of Trade Patterns And Adverse Selection, *MIS Quarterly* 33(2), 263-292.

GRIFFITHS T and STEYVERS M (2004) Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 101, 5228-5235.

HALEY DT, THOMAS P, DE ROECK A and PETRE M (2007) Tuning an LSA-based assessment system for short answers in the domain of computer science: the elusive optimum dimension. In *Mini-Proceedings of the 1st European Workshop on Latent Semantic Analysis in Technology-Enhanced Learning* (WILD F, Kalz M, van Bruggen J and Koper R, Eds), 22-23, Heerlen, NL.

HAN J and KAMBER M (2006) *Data Mining: Concepts and Techniques*, 2nd Ed. Morgan Kaufmann (Elsevier), San Francisco.

HOVORKA D and LARSEN K (2006) Enabling agile adoption practices through network organizations. *European Journal of Information Systems* 15(2), 159-168.

HOVORKA D, LARSEN K and MONARCHI D (2009) Conceptual Convergences: Positioning Information Systems among the Business Disciplines. In *Proceedings of the 17th European Conference on Information Systems (ECIS)* (NEWELL S, WHITLEY E, POULOUDI N, WAREHAM J and MATHIASSEN L Eds), manuscript 0217.R1, published by Università di Verona and London School of Economics.

HU X, CAI Z, WIEMER-HASTINGS P, GRAESSER AC AND MCNAMARA DS (2007) Strengths, Limitations, and Extensions of LSA. In *Handbook of Latent Semantic Analysis* (LANDAUER TK, MCNAMARA DS, DENNIS S and KINTSCH W, Eds), pp 401-425, Lawrence Erlbaum Associates, Mahwah, New Jersey.

HUSBANDS P, SIMON H and DING CH (2001) On the Use of the Singular Value Decomposition for Text Retrieval. In Computational Information Retrieval, (BERRY M Ed), pp 145-156, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

JOHNSON RA and WICHERN DW (2007) *Applied Multivariate Statistical Analysis*. Pearson/Prentice Hall, New Jersey.

KUECHLER WL (2007) Business Applications of Unstructured Text. *Communications of the ACM*, 50(10), 86-93.

LANDAUER TK (2007) LSA as a Theory of Meaning. In *Handbook of Latent Semantic Analysis* (LANDAUER TK, MCNAMARA DS, DENNIS S and KINTSCH W, Eds), pp 3-32, Lawrence Erlbaum Associates, Mahwah, New Jersey.

LANDAUER T, LAHAM D and DERR M (2004) From Paragraph to Graph: Latent Semantic Analysis for Information Visualization. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 101, 5214-5219.

LARSEN KR, MONARCHI DE, HOVORKA DS and BAILEY CN (2008) Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems. *Decision Support Systems* 45, 884-896.

LARSEN KR and MONARCHI DE (2004) A Mathematical Approach to Categorization and Labeling of Qualitative Data: the Latent Categorization Method. *Sociological Methodology* 34(1), 349-392.

LIFCHITZ A, JHEAN-LAROSE S and DENHIÈRE G (2009) Effect of tuned parameters on an LSA multiple choice questions answering model. *Behavior Research Methods* 41(4), 1201-1209.

MARTIN and BERRY M (2007) Mathematical Foundations Behind Latent Semantic Analysis. In *Handbook of Latent Semantic Analysis* (LANDAUER TK, MCNAMARA DS, DENNIS S and KINTSCH W, Eds), pp 33-57, Lawrence Erlbaum Associates, Mahwah, New Jersey.

MANNING C, RAGHAVAN P and SCHÜTZE H (2008) *Introduction to Information Retrieval*. Cambridge University Press, New York.

MEI Q and ZHAI C (2005) Discovering Evolutionary Theme Patterns from Text – an Exploration of Temporal Text Mining. In *Proceedings of the Eleventh ACM SIGKDD* (VAIDYA J, Ed), p 189, ACM Publications, Baltimore, Maryland.

MEROÑO-CERDAN AL AND SOTO-ACOSTA P (2007) External Web content and its influence on organizational performance. *European Journal of Information Systems* 16(1), 66-80.

MOORE GC AND BENBASAT I (1991) Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research* 2(3), 192-222.

MORINAGA S and YAMANISHI K (2004) Tracking Dynamics of Topic Trends Using a Finite Mixture Model. In *Proceedings of the Tenth ACM SIGKDD* (JOYDEEP G, Ed), 811-816, ACM Publications, Baltimore, Maryland.

O'DONOGHUE PG and MURPHY MH (1996) Object modelling and formal specification during real-time system development. *Journal of Network and Computer Applications* 19(4), 335-352.

ORD T, MARTINS E, THAKUR S, MANE K and BÖRNER K (2005) Trends in Animal Behaviour Research (1968-2002): Ethoinformatics and the Mining of Library Databases. *Animal Behaviour*, 69, 1399-1413.

PALVIA P, LEARY D, MAO E, MIDHA V, PINJANI P and SALAM AF (2004) Research methodologies in MIS: an update. *Communications of the AIS* 14, article 24.

PANTELI A, STACK J, ATKINSON M and RAMSAY H (1999) The status of women in the UK IT industry: an empirical study. *European Journal of Information Systems* 8(3), 170-182.

PARK L and RAMAMOHANARAO K (2009) An Analysis of Latent Semantic Term Self-Correlation. *ACM Transactions on Information Systems*, 27(2), 8:1-8:35.

PENUMATSA P, VENTURA M, GRAESSER AC, LOUWERSE M, HU X, CAI Z and FRANCESCHETTI DR (2006) The right threshold value: what is the right threshold of cosine measure when using Latent Semantic Analysis for evaluating student answers? *International Journal on Artificial Intelligence Tools* 15(5), 767-777.

PORTER M (1980) An Algorithm for Suffix Stripping. *Program* 14(3), 130-37. Republished as: PORTER M (2006) An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems* 40(3), 211-218.

POTTENGER W and YANG T (2001) Detecting Emerging Concepts in Textual Data Mining. In *Computational Information Retrieval*, (BERRY M, Ed), pp 89-106, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

SALTON G (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613-620.

SALTON G and BUCKLEY C (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*, 24, 513-523.

SHAHNAZ F, BERRY MW, PAUCA VP and PLEMMONS RJ (2006) Document clustering using nonnegative matrix factorization. *Information Processing and Management* 42, 373-386.

SIDOROVA A, EVANGELOPOULOS N, VALACICH JS, and RAMAKRISHNAN T (2008) Uncovering the Intellectual Core of the Information Systems Discipline. *MIS Quarterly* 32(3), 467-482 & A1-A20.

SPOMER JE (2009) Latent Semantic Analysis and Classification Modeling in Applications for Social Movement Theory. MS Thesis, Department of Mathematical Sciences, Central Connecticut State University.

STEYVERS M and GRIFFITHS T (2007) Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis* (LANDAUER TK, MCNAMARA DS, DENNIS S and KINTSCH W, Eds), pp 427-448, Lawrence Erlbaum Associates, Mahwah, New Jersey.

TEH YW, JORDAN MI, BEAL MJ and BLEI DM (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101, 1566-1581.

VALLE-LISBOA JC and MIZRAJI E (2007) The uncovering of hidden structures by Latent Semantic Analysis. *Information Sciences* 177(19), 4122-4147.

WEBER RP (1990) *Basic Content Analysis*, 2nd ed. Newbury Park, CA.

WEI C-P, YANG CC and LIN C-M (2008a) A Latent Semantic Indexing-based approach to multilingual document clustering. *Decision Support Systems* 45, 606-620.

WEI, C-P, HU P. J-H, TAI C-H, HUANG C-N AND YANG C-S (2008b) Managing Word Mismatch Problems in Information Retrieval: A Topic-Based Query Expansion Approach. *Journal of Management Information Systems* 24(3), 269-295.

WILLCOCKS L, WHITLEY EA and AVGEROU C (2008) The ranking of top IS journals: a perspective from the London School of Economics. *European Journal of Information Systems* 17(2), 163-168.

WITTEN IH and FRANK E (2005) *Data Mining: practical machine learning tools and techniques*, 2nd Ed. Morgan Kaufmann, San Francisco, California.

ZHU M and GHODSI A (2006) Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis* 51(2), 918-930.

Appendix

Table A1 Selected LSA software solutions

Software	Text Pre-processing	Core LSA (SVD)	Post-LSA extensions	*URL*
MC (in C++)	x			http://userweb.cs.utexas.edu/users/dml/software/mc/
TMG (in Matlab)	x			http://scgroup20.ceid.upatras.gr:8000/tmg/
WordNet	x			http://wordnet.princeton.edu/wordnet/download/
JAMA (in Java)		x		http://math.nist.gov/javanumerics/jama/
SenseClusters	x	x		http://senseclusters.sourceforge.net/
SVDPACK	x	x		http://www.netlib.org/svdpack/
SVDLIBC	x	x		http://tedlab.mit.edu/~dr/SVDLIBC/
Matlab™		x	x	Commercial product by MathWorks
Mathematica®		x	x	Commercial product by Wolfram Research
SAS®		x	x	Commercial product by SAS Institute
CLUTO	x	x	x	http://glaros.dtc.umn.edu/gkhome/views/cluto
Infomap	x	x	x	http://infomap-nlp.sourceforge.net/
LPU	x	x	x	http://www.cs.uic.edu/~liub/LPU/LPU-download.html
LSA	x	x	x	http://lsa.colorado.edu/
LSI BY Telcordia	x	x	x	http://lsi.research.telcordia.com/
phplsa	x	x	x	http://sourceforge.net/projects/phplsa/
SAS® Text Miner™	x	x	x	Commercial product by SAS Institute

Post-publication additions (updated April 2013):

R/LSA	http://cran.r-project.org/web/packages/lsa/index.html
SAS ® Text Miner 12.1	http://support.sas.com/software/products/txtminer/