# Discovering latent topical structure by second-order similarity analysis

Timothy Cribbin

Department of Information Systems and Computing,

Brunel University,

Uxbridge. UK. UB8 3PH.

Email: timothy.cribbin@brunel.ac.uk

## Abstract

Document similarity models are typically derived from a (word) term-document vector space representation by comparing all vector-pairs using some similarity measure. Computing similarity directly from a 'bag of words' model can be problematic because term independence causes the relationships between synonymous and related terms and the contextual influences that determine the 'sense' of polysemous terms to be ignored. This paper compares two methods that potentially address these problems by modelling the higher-order relationships that lie latent within the original vector space. The first is latent semantic analysis (LSA), a dimension reduction method which is a well known means of addressing the vocabulary mismatch problem in information retrieval systems. The second is the lesser known, yet conceptually simple approach of second-order similarity (SOS) analysis, where similarity is measured in terms of profiles of first-order similarities as computed directly from the term-document space. Nearest neighbour tests show that SOS analysis produces similarity models that are consistently better than both first-order and LSA derived models at resolving both coarse and fine level semantic clusters. SOS analysis has been criticised for its cubic complexity. A second contribution is the novel application of vector truncation to reduce the run-time by a constant factor. Speed-ups of four to ten times are found to be easily achievable without losing the structural benefits associated with SOS analysis.

**Keywords:** document similarity; vector space model; second-order similarity; latent semantic analysis; singular value decomposition; vector truncation

## Introduction

A document similarity model (DSM) is an undirected graph or matrix describing the mutual similarities of all document pairs within a corpus. Inter-document similarities are often derived automatically from the term-document vector space representation. The so-called 'bag of words' approach has a long history within information retrieval (van Rijsbergen, 1979; Salton and McGill, 1983) and, despite its simplicity, is known to be an effective means of generating cognitively plausible similarity models (Lee et al., 2005). DSMs have applications in many information retrieval and text mining tasks, such as in clustering or scaling tasks where the objective is to provide an ad hoc semantic organisation of documents to support analysis and exploration (Chen, 1998; Hornbæk, and Frokjær, 1999; Cribbin, 2010).

Whilst the bag of words approach is powerful, the inherently high-dimensionality and sparseness of the vector space, in combination with the effects of feature independence and vocabulary mismatch conspire to limit the structural validity of computed similarities. In this paper we compare two different approaches to improving the structure of the DSM: Latent Semantic Analysis (LSA: Deerwester et al., 1990) and second-order similarity (SOS) analysis (Janssens, 2007). Both methods are promising due to their ability to infer higher-order (transitive) relationships that lie latent within the structure of the term-document matrix. However, we posit that SOS analysis is both a more simple and reliable method of resolving the inherent limitations of the vector space model.

LSA has been extensively applied to the document clustering problem as a means of improving both the efficiency (Schütze and Silverstein, 1997; Mecca et al., 2007) and quality (Ahlgren and Colliander, 2009) of the clustering process. Central to LSA is the algebraic technique of singular value decomposition (SVD), a two-mode form of factor analysis, which attempts to explain, in decreasing quantities, the variance present in the term-document matrix. This method can reduce the complexity of the vector space from several thousand dimensions, down to just twenty or so factors (concepts) with no resulting loss in clustering performance (Schütze and Silverstein, 1997). However, whilst small improvements in classification accuracy have been observed (Ando, 2000; Ahlgren and Colliander, 2009), any accuracy gain achieved from LSA seems highly dependent upon the selection of the correct number of top-ranking factors (Deerwester et al., 1990), which tends to vary from one dataset to another. Whilst many heuristics exist, a simple, reliable, automatic method of doing this remains elusive (Janssens, 2007).

Second-order similarity (SOS) analysis, on the other hand, suffers no such parameter problem and evidence has started to appear in the literature to suggest it may even be a more effective method than LSA (Janssens, 2007; Ahlgren and Colliander, 2009). However, it is difficult to draw definitive conclusions from these studies as they were both restricted to single datasets and measured cluster quality in terms of post-clustering classification accuracy. In this paper, we provide a detailed evaluation of the performance of both LSA and SOS analysis using six different corpora. Moreover, cluster quality is measured within the DSM itself, using an adapted version of the nearest-neighbour test (Voorhees, 1985), thus removing any potentially confounding effect of clustering/scaling algorithm.

Despite its promise, the high computational demands of second-order similarity analysis - $O(N^3)$ – may prohibit its use in real-time or large-scale modelling tasks (Janssens, 2007). To address this issue the novel method of applying vector truncation to the FOS matrix is proposed as a means of reducing the complexity of the procedure. It was expected that truncated SOS analysis should produce similar and potentially better DSMs than the full-vector approach, given the tendency for all but the most salient (nearest neighbour) FOS scores to skew towards zero, a result of the sparseness inherent to the term-document space (Martin-Merino and Munoz, 2004).

The rest of this paper is organised as follows. In the next section, the motivation for this work is explained in greater detail along with formal definitions of the second-order similarity and other methods applied in this study. Following a description of the methodology used, the results of a series of experiments, examining cluster quality and run-time performance, are presented and discussed. In the final section, conclusions are drawn and implications for future work discussed.

## Discovering latent semantic structure

Perhaps the most popular method used to compute inter-document similarity is the vector space model (VSM: Salton and McGill, 1983). In the VSM, documents are first encoded as points within a common, high-dimensional term space. The simplest encoding is the 'bag of words' (BOW) approach, wherein term dimensions are defined as unique word tokens occurring within the corpus. A vector is constructed for each document by assigning a non-negative weight along each dimension based the local frequency of the term. This can be either binary (present or absent) or a frequency count (See Lee et al., 2005). Weights are often then adjusted using some kind of global scheme (e.g. inverse document frequency) in order to increase the relative influence of rarer, more discriminating terms. The resulting vector space can be represented formally as an *M x N* matrix, where *M* is the number of unique terms within the corpus and *N* is the number of documents.

Document similarities are then computed from this space, either as the distance between document vector points or, more commonly, as the angle between points relative to the origin. In this paper we adopt the latter approach, computing the cosine (normalised dot product) between all document vector pairs. More formally the *N x N* dimensional FOS matrix, $\boldsymbol{B}$, is derived as the normalised product of the transposed *M x N* dimensional term-document matrix, $\boldsymbol{A}^{\boldsymbol{T}}$, with the original matrix, $\boldsymbol{A}$ i.e.:

$$\boldsymbol{B_{BOW} = A^{T}.A}$$

Given appropriate term selection and weighting, the BOW model can produce cognitively plausible results (Westerman et al., 2010; Lee et al., 2005). However, the validity of resultant similarities can be compromised by *vocabulary mismatch* which is the tendency for different words to be used to express the same concepts (Furnas et al., 1987). As terms are deemed independent, the semantic relationship between documents using synonymous or semantically related terms will remain obscured, wrongly suppressing similarity measures. Likewise, words can often have multiple senses, depending on the context in which they are used. For example, "bank" means different things in a document about finance, as opposed to one about fishing, yet the term will be represented as only one dimension in the term-document matrix. Such polysemous words can erroneously inflate the measured similarity between semantically unrelated documents.

### *Latent semantic analysis*

A popular solution to the vocabulary mismatch problem is to transform the BOW model using a procedure called latent semantic analysis (LSA: Deerwester et al., 1990). LSA transforms the BOW matrix into a new space where both terms and documents are represented in terms of a smaller number of statistically-derived factors or *concepts*. LSA is achieved by means of a two-mode factor-analysis technique known as singular value decomposition (SVD). Formally, after SVD, the term-document matrix, ***A***, is decomposed to three new matrices such that:

$$\boldsymbol{A \rightarrow T.S.D^{T}}$$

***T*** is an $M$ *x* $R$ dimensional term-concept matrix, ***D*** is an $N$ *x* $R$ dimensional document-concept matrix and ***S*** is an $R$ *x* $R$ diagonal matrix containing singular values. Singular values can be thought of as weights or scaling factors that describe the relative global importance of each concept. The number of concepts, $R$, required to preserve all of the original variance is min($M, N$). This is usually $N$ as the number of terms, $M$, generally exceeds the number of documents in all but very large corpora. As 100% of the original variance is preserved, it is possible to reconstruct the original term-document matrix from the derived matrices.

Concepts are orthogonal (uncorrelated) factors, meaning that they each account for separate portions of the original variance in ***A***, and are ranked in order of decreasing singular value (variance accounted for). The distribution of singular values tends to follow a Zipfian distribution, with the higher-ranking concepts accounting for relatively larger amounts of the original variance. This property makes it possible to discard most of the lower ranking dimensions without sacrificing the salient semantic structure, a process known as *rank reduction* or sometimes *global truncation* (Schutze and Silverstein, 1997). Note that for the remainder of this paper, we use the term rank reduction when referring to truncation of an LSA document-concept space and the term global truncation is used to refer to one of the proposed methods (explained later) of truncating the FOS matrix prior to SOS analysis.

Computing the document similarity matrix from LSA space is much the same as the procedure for the original term-document space. In essence the matrix ***B*** is derived by multiplying ***D*** by its transpose i.e.:

$\boldsymbol{B_{LSA} = D^T.D}$

However, it is also usual practice to scale the singular vectors, ***D* and *D*$^T$**, prior to computing similarity, by multiplying each vector element by its corresponding singular value (Deerwester et al., 1990):

$\boldsymbol{D_{ic} = D_{ic}.S_{c,c}}$

Given that full-rank SVD preserves all of the original variance, when the rank of ***D*** is equal to *R* (no rank-reduction), the derived matrix $\boldsymbol{B_{LSA}}$ will be identical to $\boldsymbol{B_{BOW}}$. However, it is typical in clustering applications to employ rank-reduction of SVD space as a means of reducing computation time. Empirical studies have found that even very aggressive rank reduction, down to rank-50 or even rank-20 has little or no significant impact on the resulting cluster validity (Schutze and Silverstein, 1997; Janssens, 2007; Mecca et al., 2007). As the number of unique terms defining a document can be quite close to its total word count, particularly in the case of shorter documents, reducing dimensionality to this extent reduces the computation time for the similarity matrix quite considerably over using the original BOW term-document space assuming that the documents are already indexed as an SVD space (Schutze and Silverstein, 1997).

A more interesting question, however, is the potential for LSA to *improve* the semantic validity of inter-document similarities. LSA has been shown to be a powerful learning algorithm that can effectively model second- order semantic relations between terms without access to any pre-defined semantic network (Deerwester et al., 1990; Kontostathisa and Pottenger, 2006). For example, “football” and “soccer” are synonyms so rarely occur together in the same document. However, their semantic association can be modelled because they tend to co-occur with the same related terms (e.g. goal, penalty, defender etc.). After SVD, both football and soccer are likely to be associated with the same factor(s). Likewise, LSA may help to alleviate the polysemy problem caused when vocabulary terms possess multiple senses. Polysemy can erroneously increase similarity between unrelated documents. For example, the word “bank” can be used within contexts as diverse as river geography and financial systems.  Following SVD, “bank” would load onto (be associated with) separate factors i.e. one comprised mostly of financial terms and another composed of terms relating to rivers. Within ***D***, a financial article would likely have a large weight on the former factor and only a relatively small weight on the latter factor, whilst an article on fishing would exhibit the reverse pattern in its singular vector.

For these reasons, LSA has proved extremely successful in both information retrieval and semantic reasoning applications. By transforming query terms into a singular vector, it is possible for the system to return a relevant document even if the document contains none of the words specified by the user (Deerwester et al, 1990).  This is possible because synonymous and other, closely related words will tend to load on the same factors. Moreover, because most of the variance is modelled in the top factors, even very large corpora can be effectively indexed in just 150 to 300 dimensions, improving retrieval speed without negatively affecting quality (Deerwester et al., 1990). Latent semantic models also make it possible to automate various other language tasks that are impossible using a bag of words VSM. For instance, Landaur and Dumais (1994) found that LSA could achieve the same score in a standardised multiple choice language test – identifying synonyms – as the average student taking the test. It has even been shown that is possible to use LSA models to grade student essays, with an accuracy that approaches that of human judges (Landauer et al., 1997).

Whilst there is little doubt that LSA effectively models second and higher order semantic relations between terms (Kontostathisa and Pottenger, 2006), the concrete advantage LSA affords to inter-document similarity modelling is less certain. As factors are ranked according to decreasing variance accounted for, it seems reasonable to assume that eliminating lowest ranked factors from the space may also help to reduce the impact of minor, random variation caused, for instance, by typographical errors and use of non-discriminating terms. Ahlgren and Colliander (2009) found that a reduced rank (top 90% of cumulative singular values) SVD document space improved cluster classification accuracy of a small (43 documents) corpus by around 19%, as measured by the Rand index (Rand, 1971) as referenced to a ground-truth classification of a domain expert. In contrast, Ando (2001) found less impressive gains when studying a larger sample of corpora. Using the optimal rank reduction for each corpus, determined manually (average rank reduction was just under 40%), the improvement in average precision of intra-topic pairs was just 3% over that achieved using the original term-document vectors.

A key difficulty with rank reduction is that of selecting the optimal number of factors to retain (Janssens, 2007) as this can vary widely from one similarly sized corpus to the next (Ando, 2000). One common heuristic is to select the elbow point from the scree plot of singular values so as to retain the most variance for the least number of dimensions. A more easily automated heuristic is to select factors down to the average singular value. However, there is little agreement between users of LSA. Some suggest that fewer factors are most likely to lead to optimal clustering solutions (e.g. Janssens, 2007) whilst others have advocated a strategy in which most (e.g. 90%) but not all of the original variance is retained (Ahlgren and Colliander, 2009). To confound this uncertainty, He et al. (2004) suggest that rank-reduced LSA may be fundamentally unsuitable for modelling document similarity because, being a global method, SVD extracts the most thematic (relating) features first, rather than the finer, discriminatory features that are key to detecting the more subtle differences and associations between documents (Salton et al., 1975). We argue that an effective DSM should discriminate between clusters of documents at both higher and lower levels of semantic granularity. In rejecting lower ranked factors there is the inherent danger of losing sensitivity to minor or distinct themes (Ando, 2000) that may be crucial when exploring a complex topic (Muresan and Harper, 2004) as a direct relationship between singular value and information value (signal-noise) cannot be guaranteed for many datasets (Aggarwal, 2001). Given this, along with the computational complexity of LSA – $O(\min(MN^2, M^2N))$ – coupled with the difficulty associated with determining the optimal factors, a simpler and more reliable method of determining a more accurate DSM is desirable.

### *Second-order similarity analysis*

Second order similarity (SOS) analysis offers an alternative method of extracting the latent structure from a VSM that may be more suited to document similarity modelling (Janssens, 2007; Ahlgren and Colliander, 2009). Rather than attempting to model the semantic relationships between terms, per se, the technique focuses on the document as the unit of analysis, taking the initial matrix of FOS coefficients and re-computing similarity between each document pair as the angle or correlation between their similarity profiles. More formally, the SOS matrix, ***C***, is derived as the normalised product of the FOS matrix ***B*** with its transpose i.e.:

$\boldsymbol{C = B.B^T.}$

This technique is closely related to the bibliometric analysis technique commonly used to transform co-citation (e.g. author, reference) frequency data into a common metric scale. McCain (1990) noted that correlating co-citation profiles has the added advantage of inferring similarity through transitivity i.e. if author/document A and author/document B are both regularly co-cited with author/document C, then it is reasonable to assume that A and B are also related in some way (McCain, 1990). Likewise, it is reasonable to assume that documents sharing a similar FOS profiles are also likely to be related, even if their direct FOS is relatively low. In other words, second-order similarity may resolve problems of vocabulary mismatch, allowing latent semantic relations to be effectively identified. Following a comprehensive comparison of similarity metrics against a benchmark of 'gold standard' human similarity judgements, Lee et al. (2005) concluded that a similarity model that "judges documents in terms of their similarity to other documents" may be a more cognitively plausible approach.

The logic behind SOS analysis can be demonstrated with a simple example. Let there be four documents, D1 to D4. D1 and D2 are highly similar, whilst D4 is an outlier that is only weakly similar to the other documents. D3, on the other hand is highly similar to D2 but only weakly similar to D1 and D4. Figure 1 shows the FOS matrix on the left-hand side and the resulting SOS matrix on the right. Whilst all similarity coefficients increase to some extent, note how SOS analysis correctly makes the inference that D3 is more likely to be similar to D1 (0.09 → 0.29) than to D4 (0.09 → 0.20). This inference is made on the basis that both D1 and D3 are close neighbours of D2, unlike D4, which is only weakly related to the salient cluster.

| D1 | D2 | D3 | D4 | | D1 | D2 | D3 | D4 |
|---|---|---|---|---|---|---|---|---|
| 1.00 | 0.42 | 0.09 | 0.06 | | 1.00 | 0.69 | **0.29** | 0.15 |
| 0.42 | 1.00 | 0.38 | 0.11 | → | 0.69 | 1.00 | 0.65 | 0.24 |
| 0.09 | 0.38 | 1.00 | 0.09 | → | 0.29 | 0.65 | 1.00 | 0.20 |
| 0.06 | 0.11 | 0.09 | 1.00 | | 0.15 | 0.24 | **0.20** | 1.00 |

Figure 1: SOS transformation of a simple FOS matrix

Bear in mind that this is a very simple example and that adjustments made by SOS analysis will typically be more subtle, being influenced by many more FOS data points. *Our first hypothesis (H1) is, therefore, that second-order similarity analysis can reliably detect latent semantic structure that is invisible to FOS analysis.*

Two recent studies have provided empirical evidence to support this hypothesis (Janssens, 2007; Ahlgren and Colliander, 2009). Ahlgren and Colliander (2009), for instance, compared the quality of outputs from a clustering algorithm for a set of 43 articles retrieved from *Information Retrieval*. Articles were initially classified by an expert in the field to provide the ground-truth benchmark. FOS DSMs were then created using both word and LSA singular vectors (retaining 90% of cumulative singular values) and second-order DSMs then derived from both of these matrices. They found that whilst LSA improved clustering, particularly after SOS analysis, the second-order matrix derived from the FOS matrix resulted in by far the best classification performance overall.

Janssens (2007) had found positive results prior to this study, using larger document sets, but only measured performance in terms of intra-cluster coherence, using the silhouette measure. The author rejected SOS as a candidate for further analysis due to the perceived scalability (both space and time) problems, although they acknowledged that the technique is a viable alternative to LSA for small to moderately sized corpora (i.e. <10,000 documents).

### *Truncated second-order similarity analysis*

We have already discussed the use of vector truncation to improve clustering performance in SVD spaces (e.g. Schutze and Silverstein, 1997). Truncation has also been shown to be an effective measure when using term-document spaces (see Larsen and Aone, 1999). However, to our knowledge, truncation has never been applied as a means of improving the efficiency of SOS analysis. The rationale for its application to FOS analysis is based on the assumption that the majority of non-zero term-document element values carry little semantic value, representing only noise caused by incidental word use (Madsen et al, 2004). When local and global weighting schemes (e.g. TFIDF) are applied it is possible to rank vector elements by their informational salience. Results show that even highly aggressive pruning down to the top fifty weighted terms (and lower), has little impact on the validity of clustering solutions (Schutze and Silverstein, 1997; Larsen and Aone, 1999).

It is therefore proposed that the same logic can be applied to SOS analysis. Empirical evidence shows that within a typical term-document vector space, most document pairs are not close neighbours at all; in fact the majority of pairs will tend to be almost equally distal, with the peak of a typical similarity distribution being strongly skewed towards zero (Cribbin, 2010; Muresan and Harper, 2004; Martín-Merino and Muñoz, 2004). This phenomenon is often referred to as the *curse of dimensionality*. True regions of semantic association (i.e. clusters of topically related items) are most likely to be represented by the relatively small proportion of values that fall within the long tail beyond the mode, as predicted by the cluster hypothesis and confirmed, for example, by Muresan and Harper's (2004) empirical study. Hence, computing second-order similarities using only the top-ranked FOSs should theoretically have little impact on the quality of second-order similarities and may even improve them if the truncation policy is effective in reducing more noise than information.

Although truncation does not reduce the growth rate of SOS analysis in relation to corpus size - complexity remains O($N^3$) – the speed-up should be at least proportional to the average degree of truncation per document vector. In this paper, we use the term *K* to describe the mean percentage of non-zero elements retained per document vector. The expected speed-up would therefore be inversely proportional to K (i.e. 100/K). In other words K=25 should yield a speed-up of some four times over full vector SOS analysis.

We begin our experimentation with two alternative truncation schemes. The first, *local truncation*, is a democratic scheme in which the top K percent of elements are retained in every document vector and the remaining elements are zeroed. This is equivalent to the local truncation method proposed by Schutze and Silverstein (1997) to improve the efficiency of FOS analysis. The second approach is *global truncation*, whereby a single threshold is applied to the entire FOS matrix. The global threshold is established as the top $K^{th}$ percentile of the entire similarity distribution and all elements with values below this threshold are zeroed. Thus, in this scheme, the truncation rate will vary from one document vector to another.

We have described how the most semantically salient relationships fall into the long upper tail of the similarity distribution. Global truncation exploits this phenomenon because it preserves relatively more elements within the vectors of more central documents, which tend to have a greater number of true neighbours than more outlying documents. In contrast, a local truncation scheme assumes that all documents reside in the centre of regions of equal density, which runs the risk of salient similarity values being pruned from the central vectors, on the one hand, and semantically insignificant values being retained within the vectors of outlying documents on the other. Based on this reasoning, the next hypothesis (H2) tested in this paper is that *global truncation will yield, for any given mean truncation rate, better results than the local scheme.*

**Local Truncation**

| D1 | D2 | D3 | D4 |
|---|---|---|---|
| **1.00** | **0.42** | 0.09 | 0.06 |
| **0.42** | **1.00** | **0.38** | **0.11** |
| 0.09 | 0.38 | **1.00** | 0.09 |
| 0.06 | 0.11 | 0.09 | **1.00** |

→

→

**Resulting SOS matrix**

| D1 | D2 | D3 | D4 |
|---|---|---|---|
| 1.00 | 0.71 | 0.14 | 0.04 |
| 0.71 | 1.00 | 0.33 | 0.10 |
| 0.14 | 0.33 | 1.00 | 0.04 |
| 0.04 | 0.10 | 0.04 | 1.00 |

**Global Truncation**

| D1 | D2 | D3 | D4 |
|---|---|---|---|
| **1.00** | **0.42** | 0.09 | 0.06 |
| **0.42** | **1.00** | **0.38** | 0.11 |
| 0.09 | **0.38** | **1.00** | 0.09 |
| 0.06 | 0.11 | 0.09 | **1.00** |

→

→

**Resulting SOS matrix**

| D1 | D2 | D3 | D4 |
|---|---|---|---|
| 1.00 | 0.67 | 0.14 | 0.00 |
| 0.67 | 1.00 | 0.62 | 0.00 |
| 0.14 | 0.62 | 1.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 1.00 |

Figure 2: SOS transformation of a simple matrix following local and global truncation of the FOS matrix

Figure 2 shows a comparison of the local and global schemes using the same simple case as before. It can be seen that the global method more strongly emphasises the differences between D4 and the main D1, D2 cluster, whilst simultaneously pushing D3 closer towards the main cluster. This example also highlights a potential disadvantage of global truncation, which is a risk of the SOS graph becoming disconnected. However, as the results presented later show, for real datasets of non-trivial size, truncation must be particularly aggressive (K << 25%) to cause this.

Another disadvantage of global truncation is that it is relatively expensive to compute. Computing the exact local threshold has a complexity of $O(nN^2)$, where n is the number of elements retained (i.e. KN/100). In contrast finding the exact global threshold has a complexity of $O(n^2N^2)$. Later in this paper, we propose two more efficient methods of that approximate the global threshold with acceptable accuracy, but in much lower time. Specifically, we propose one method that estimates the global threshold as the mean of a sample of local thresholds and a heuristic method which assigns the global mean as the threshold. Application of both of these approaches yields overall speed-ups that are close to theoretical expectations.

### *Summary*

In summary, the validity of clusters resolved within an FOS matrix derived from a BOW representation can be compromised by problems of vocabulary mismatch. LSA can theoretically resolve this problem, but the empirical evidence is inconclusive and complicated by uncertainty over how to set the threshold for rank-reduction. In the first part of the experiment we examine five different rank-reduction schemes that span a wide range of original variance accounted for (50% - 97%). We expected that LSA would produce at least some improvement over FOS, but that the optimal rank threshold would vary somewhat from case to case. On the other-hand, following Ahlgren and Colliander (2009), we expected that full vector SOS analysis would reliably achieve improvements in cluster validity that were at least as good as those achieved by optimal LSA. Our intention was to generalise Ahlgren and Colliander's findings by comparing the performance of FOS, LSA and SOS across six different and much larger corpora.

In the second part of experiment, we test the proposed method of vector truncation as a means of reducing the run-time overhead of SOS analysis. Truncation was manipulated from an average of 50% down to just 5% of the original number of elements (N) in each FOS vector. Cluster validity was compared to that of full-vector SOS, using both global and local truncation schemes. The expectation was that global truncation would be the more effective scheme.

In the third part of the experiment, run-time performance is evaluated using corpora ranging in size from N=100 to N=2235.  Finally, two-dimensional spatial projections are presented to provide a tangible appreciation of the structural improvements that can achieved by SOS analysis.

## Method

This section first describes the datasets and text analysis methods used to create the DSMs. It then describes the measurement and analysis methods used to perform the evaluations. Unless stated otherwise, all software used in these experiments was designed and implemented by the author using Microsoft Visual Basic 6. All run-time measures were collected using compiled code executed as a single thread on an AMD Athlon 7750 (2.7GHz) processor and 3.25GB of available RAM.

### *Document corpora*

The corpora used in the main cluster validation experiments were composed of documents from the Financial Times news article archive as provided by the TREC Research Collection (Volume 4, 1996). Documents were retrieved using high-recall queries (one or two key terms) based on a sample of six topic descriptions taken from the TREC interactive tracks 6, 7 and 8 (Voorhees and Harman, 1997-1999). The interactive track was a competition where participants (users) attempted to find as many distinct answers to questions defined by given topic descriptions (Voorhees and Harman, 1997). After the competition, the top results retrieved by each participant were concatenated to form a pool of relevant document candidates. The TREC assessor then examined each document in turn, gradually compiling a list of different 'aspects' of relevance. Once an exhaustive list of aspect definitions was formed, all documents were then classified as relevant or not relevant to each aspect. As such, relevance judgments form a two-level hierarchy whereby all pooled documents are either relevant or not relevant to each aspect, with all documents that are relevant to at least one aspect being, by definition, relevant to the general topic.

Six topics were selected, two from each of three conferences (TREC 6, 7 and 8) so as to minimise potential bias in topic definitions and relevance judgments. Topic selection was quite arbitrary, the only requirement being a tendency for multiple documents to relate to each aspect, as defined by the relevance data, to make it possible to measure clustering tendency at this level. The final selection of topics is shown in Table 1 and included

T307i, T347i, T352i, T387i, T408i and T446i. Corpus sizes ranged from N=122 to N=588. Topic level recall was above 70% in all of the corpora. Only documents with one or more aspect relations (i.e. were not solely a member of a singleton aspect cluster) were considered in the aspect level performance analysis, hence the lower document totals. It is also important to note that whilst topic membership is exclusive, documents can, and often are relevant to more than one aspect definition. This tendency for aspect clusters to overlap is indicated in the column entitled "Mean aspects per document". It can be seen that the degree of overlap varies considerably between corpora. For instance, T307 is quite diverse (21 aspects), with all aspects clusters being relatively small (few aspect relations per document) and distinct (just one aspect per document). In contrast T387, although similar to T307 in terms of both corpus and topic sub-set size, comprises just 9 aspects clusters that tend to overlap to a great extent – each relevant document has, on average, around 70% of all other relevant documents as a same-aspect relation.

| Topic | # of documents in corpus | # documents in topic cluster | # aspects | Mean aspects per document | # documents in aspect clusters | Mean aspect relations per document |
|---|---|---|---|---|---|---|
| All | | | | 1.42 | | 16.76 |
| T307 | 137 | 48 | 21 | 1.00 | 37 | 3.30 |
| T347 | 127 | 33 | 22 | 1.21 | 24 | 2.25 |
| T352 | 218 | 87 | 28 | 1.82 | 85 | 14.54 |
| T387 | 162 | 39 | 9 | 1.59 | 38 | 26.95 |
| T408 | 122 | 53 | 15 | 1.58 | 45 | 35.51 |
| T446 | 588 | 55 | 15 | 1.02 | 49 | 12.73 |

Table 1: Summary of key structural characteristics of the topics/corpora studied

In order to explore the scalability of truncated SOS, further corpora were extracted from the TREC FT dataset and the Web of Science. The multi-topic (MT) corpora were composed of documents relevant to five randomly selected TREC topics. A base corpus of 500 documents (MT500) was initially extracted, comprising 100 documents from each topic. Four further, smaller corpora (N=100 to N=400) were then derived from this corpus by randomly sampling from each topic, maintaining equal representation from each topic.

One final, much larger corpus was also constructed. The *WOS3T* dataset, comprises Web of Science sourced article metadata (title plus abstract) for 2235 papers published recently (2005-2009) in three IEEE Transactions journals: Computers (569 papers), Education (272 papers) and Nuclear Science (1394 papers). The only objective criterion for clustering was deemed to be publication source i.e. no attempt was made to classify articles into specific topics. Whilst these journals are generally distinct in their aims and scope, being all IEEE publications, one would expect some degree of overlap in technical aspects (e.g. programming techniques/languages, mathematical modelling etc.), albeit approached from different perspectives. Hence one would expect most papers to fall in to their respective journal cluster, with a sizeable minority blurring the boundaries between these three clusters. WOS3T was only used in the run-time performance and visual comparison parts of the analysis. The objective criterion was the degree of spatial separation between the journal clusters.

*Term document matrix*

The basis for all similarity models was a BOW term-document vector space model, where the weight (*W*) of each vector element was computed as the product of raw term frequency (TF) and logarithm of inverse document frequency (IDF) (Salton and McGill, 1983). Terms were not stemmed but stop-words and both frequent (df > N×0.9) and unique (df = 1) terms were removed from the vocabulary to improve discrimination and reduce vocabulary size. The resulting term-document matrix was represented as an *M* (number of unique terms) by *N* (number of documents) matrix (two-dimensional array).

*First-order similarity analysis*

First order similarity was then computed for all unique document pairs (*i, j*) using the cosine (normalised dot-product) measure, whereby the sum of all products in the range *k* = 1 to *M* are divided by the product of the length of both participating document vectors. $W_{ik}$ represents the *k*th term in the *i*th document:

$$\text{FOS } (i,j) = \frac{\Sigma\ (W_{ik} * W_{jk})}{\sqrt{\Sigma\ (W_{ik})^2 * (W_{jk})^2}}$$

*LSA*

LSA is computed using a statistical method known as singular value decomposition (SVD), a two-mode variant of factor analysis which decomposes the TDM into three new matrices, *T*, *S* and *D*. *T* and *D* are matrices that re-express term and document vectors respectively as singular vectors within a derived sub-space. *S* is the diagonal matrix of singular values. The SVD was computed using code from the freely available ALGLIB[1], which is based on the well known LAPACK libraries.

Document singular vectors in D were scaled by the singular values in S (as recommended by Deerwester et al., 1990) i.e.:

$$\mathrm{LSA}\ (i,j) = \frac{\sum (\mathrm{D}_{ik} * \mathrm{D}_{jk} * \mathrm{S}_k^2)}{\sqrt{\sum (\mathrm{D}_{ik} * \mathrm{S}_k)^2 * (\mathrm{D}_{jk} * \mathrm{S}_k)^2}}$$

Singular values are ranked from high to low according to variance accounted for. Rank reduction was achieved by limiting *R* to the desired threshold. Five different rank-reduction criteria were applied in this study. Two conditions - *LSA-20* and *LSA-50* – were based on a simple rank threshold criterion (following Schütze and Silverstein, 1997). *LSA-elb* required manual intervention as *k* is determined by visual identification of the elbow in the distribution of singular values. *LSA-90PCV* condition reduces D to the rank at which 90 percent of original variance is still accounted for (variance accounted for by a singular value is proportional to its square). Finally, *LSA-75PC* represents a rank reduction to the 75th percentile of all singular values i.e. for a space of rank 100, the top 75 factors are retained. The average original variance explained was 97.3% with only

[1] http://www.alglib.net/matrixops/general/svd.php

small variation between corpora (96.6% - 98.7%) making LSA-75PC the least aggressive rank reduction condition.

*Second-order similarity analysis*

Full-vector second-order similarities were computed by finding the cosine similarity of all document pairs according to their FOS vector profiles, where $k = 1$ to $N$:

$$\mathrm{SOS}\,(i,j) = \frac{\Sigma\,(\mathrm{FOS}_{ik} * \mathrm{FOS}_{jk})}{\sqrt{\Sigma\,(\mathrm{FOS}_{ik})^2 * (\mathrm{FOS}_{jk})^2}}$$

In co-citation analysis, there has always been some disagreement over whether to include or exclude diagonal values in the profile similarity analysis. Both self co-citation and total citation count, the alternative measure, will tend to dwarf other elements in a profile, particularly if the author or document is particularly prominent in the field or topic (McCain, 1990). However, in the case of textual similarity measures, whilst the FOS of a document to itself is still nearly always the largest element in its profile (unless a duplicate exists in the corpus), the normalisation inherent to the cosine measure ensures that this value is always one. Whilst retaining the diagonal does mean that the contribution of rows *i* and *j* to the $\mathrm{SOS}_{\mathrm{ij}}$ cosine coefficient are comparatively large, unless the pair forms an exclusive cluster in FOS space, the contribution of mutual neighbour associations is likely to be just as if not more significant. Moreover, if *i* and *j* do in fact form a cluster with no mutual neighbours, then removing the diagonal would lead to a zero SOS coefficient, which is obviously undesirable. Hence, the diagonal values were left intact.

### *Vector truncation*

For the purpose of cluster validation, vector truncation was performed using the local and global schemes as described in the previous section. In both cases, the threshold(s) were determined using a partial selection sort, as it was only necessary to sort to the Kth percentile. For the run-time analysis, we implemented two alternative, approximate methods of determining the global threshold that successfully overcome the complexity problem outlined earlier. The first is an expedient method that sets the threshold as the mean proximity value of the FOS matrix. This resulted in average truncation levels of K=24 to K=32 across the TREC corpora. The second method is a hybrid of the local and global schemes, which provides greater control over K. In this method, local thresholds are computed for a set sample of columns within the FOS matrix and the mean average of these thresholds is then taken as an estimate of the global threshold. Usually, this provides a generally accurate estimate of the global threshold and is considerably faster than the exact method. These are explained in greater detail in the run-time analysis sub-section of the Results and Discussion.

### *Optimisation of full and truncated vector SOS analysis routines*

The simplest implementation is to represent the FOS matrix as a two-dimensional (N, N) array and then, for each $\text{SOS}_{ij}$, iterate exhaustively (1 to N) through the rows of the two columns, *i* and *j* computing the product in each case. However, computing the product of element pairs that contain at least one zero value is a wasted operation. Whilst the original FOS matrix is generally dense it naturally becomes highly sparse following truncation. Using conditions to check for non-zero values at each step transpires to be more expensive than simply computing the unnecessary product. In order to achieve the theoretical performance gains it was necessary to build an index of non-zero elements for each column of the FOS matrix. Hence, whilst the actual data remained in a square array, for each $\text{SOS}_{ij}$ calculation, only the rows, *k*, where $\text{FOS}_{ik} > 0$ were processed. The use of these indexes makes little difference to the run-time of full vector SOS analysis, but allows truncated analysis to achieve its theoretical speed-up.

### *Spatialisation*

Spatial-semantic solutions (2D projections) were computed separately from the main experimental code using the PROXSCAL function in SPSS version 15. Matrices were loaded into SPSS in dissimilarity format (where dissimilarity = 1 – similarity) and solutions found using an ordinal scaling model, no untying of tied ranks, and a simplex start configuration. All other options were left as default.

### *Experimental Design*

Hence, the independent variable was the method of transforming the FOS matrix as derived directly from the BOW model. The experimental conditions comprised transformations using LSA (*LSA-20*, *LSA-50*, *LSA-Elb*, *LSA-90PCV* and *LSA-75PC*), full vector second-order similarity (*SOS*) analysis and truncated second-order similarity analysis (SOS-50, SOS-25, SOS-15, SO-10 and SOS-5). Third-order similarity (*TOS*) analysis was also included, as a supplementary condition, to determine whether an additional iteration would lead to further advantage. Previously, Chen (2002) has shown how further iterations of similarity profile analysis cause correlation coefficients to diverge towards either one or minus one. To the author's knowledge this is the first time TOS analysis has been investigated before within the context of document similarity analysis.

In order to validate the code and procedure used to generate the LSA conditions, DSMs were also computed using full rank singular vectors (LSA-100PC). As expected, all DSMs were all identical to those computed by the BOW method.

Cluster validity was measured using an adaptation of Voorhees (1985) nearest neighbours test, as used in previously by Cribbin (2010). This test incorporates two measures: *trustworthiness* - the tendency for a document nearest neighbours to be true neighbours - and *continuity* which is a measure of cluster cohesion (See Venna and Kaski, 2006). Both measures were computed at both the topic and aspect levels of semantic association.

Trustworthiness is based on the familiar precision function. Specifically, for each (topic) relevant document, a trustworthiness (*T*) score is computed as the proportion of its five nearest neighbours to which it is known to be related (i.e. same topic or same aspect(s)). Continuity, in contrast, is derived from the recall function. Specifically, for each relevant document, continuity (C) is the proportion of all known topical/aspectual relations residing amongst its twenty nearest neighbours. Finally, the harmonic mean of these two measures was also computed to provide a composite measure, similar to the familiar F-score used in IR evaluation (van Rijsbergen, 1979).

Differences between the control (FOS) and experimental conditions were evaluated using the Wilcoxon signed ranks test and Friedman ANOVA. The data from all six corpora shown in Table 1 were aggregated resulting in sample sizes of N=315 at the topic level of association and N=278 at the aspect level. The latter sample size is smaller due to the occurrence of singleton cases that had no aspectual relations and therefore had to be excluded from that part of the analysis. A non-parametric analysis was deemed appropriate on two counts. First, the distribution of scores on both measures tends to deviate from the normal, particularly for trustworthiness where the range of unique values is restricted to five. Secondly, there was some concern that the general change in cluster structure could be obscured in the mean scores when, for instance, a small number cases experience an unusually large change in their structural location. The signed ranks test usefully provides summary information on the frequency of cases that improve, degrade or remain unchanged (positive ranks, negative ranks and ties).

The data used to compare the run-time performance of full vector and truncated SOS analysis was collected using compiled code, with all optimizations enabled in the compiler. The timer is reset at the point where the FOS matrix has been loaded into memory and stops when the last element in the SOS matrix has been calculated. Absolute times are shown only for full vector SOS (the control). The measure reported for the other conditions is a ratio, where speed-up is defined as the run-time for full vector SOS divided by the run-time for the respective truncated condition. All times reflect the mean average time following 100 iterations.

## Results and discussion

This section proceeds in four steps. First, cluster validity is compared between the LSA conditions and those of SOS and TOS. The results show clear advantages of SOS analysis over LSA at all levels of rank-reduction. Second, the effect of vector truncation is explored, comparing both local and global schemes. The results show that moderate truncation tends to lead to slight improvements over traditional full-vector SOS analysis, with a global scheme producing noticeably better results. Third, it is demonstrated that it is possible to apply a global scheme whilst still achieving the theoretical speed gains of truncated SOS analysis by applying either approximate or heuristic methods of threshold setting. In the final sub-section, both full vector and truncated analysis solutions are visualized as two-dimensional projections or spatialisations. These visualisations provide a compelling illustration of how SOS analysis can improve semantic structure.

*LSA*

Table 2 shows the mean T, C and F scores for all LSA conditions and SOS and TOS along with the percentage change relative to FOS. The LSA conditions are shown in descending order of the average original variance accounted for (% Var).

| | FOS | LSA-75PC | LSA-90PCV | LSA-50 | LSA-Elb | LSA-20 | SOS | TOS |
|---|---|---|---|---|---|---|---|---|
| % Var | | 97.3 | 90.0 | 74.5 | 56.2 | 52.6 | | |
| **F-Asp** | ***.3887*** | **.3259***<br>**-16.2%** | **.3306***<br>**-15.0%** | **.3399***<br>**-12.6%** | **.3332***<br>**-14.3%** | **.3395***<br>**-12.7%** | **.4101***<br>**+5.5%** | **.3817**<br>**-1.8%** |
| T-Asp | *.3950* | .3317*<br>-16.0% | .3496*<br>-11.5% | .3712+<br>-6.0% | .3633+<br>-8.0% | .3712+<br>-6.0% | .4216*<br>+6.7% | .4014<br>+1.6% |
| C-Asp | *.5482* | .4829*<br>-11.9% | .4662*<br>-15.0% | .4696*<br>-14.3% | .4447*<br>-18.9% | .4502*<br>-17.9% | .5703*<br>+4.0% | .5345<br>-2.5% |
| **F-Top** | ***.2779*** | **.2414***<br>**-13.1%** | **.2371***<br>**-14.7%** | **.2315***<br>**-16.7%** | **.2492***<br>**-10.3%** | **.2487***<br>**-10.5%** | **.2941***<br>**+5.8%** | **.2765**<br>**-0.5%** |
| T-Top | *.5137* | .4546+<br>-11.5% | .4597+<br>-10.5% | .4692+<br>-8.7% | .4927<br>-4.1% | .4927<br>-4.1% | .5448*<br>+6.1% | .5251<br>+2.2% |
| C-Top | *.2076* | .1839*<br>-11.4% | .1769*<br>-14.8% | .1715*<br>-17.4% | .1830*<br>-11.8% | .1822*<br>-12.2% | .2197*<br>+5.8% | .2065<br>-0.5% |

Table 2: T, C and F scores along with percentage improvements of LSA, SOS and TOS over FOS. Significant differences to FOS are highlighted (* p<.001).

Looking first at the F scores, it is clear that all LSA methods result in a general and highly significant (p<.001) loss of structural validity, at both topic and aspect levels of association, regardless of the chosen rank threshold. Non-parametric ANOVA reveals a highly significant (p<.001) general effect of LSA rank threshold, at both levels of semantic association, but particularly at the topic level.

As expected, there appears to be no simple linear relationship between the degree of rank reduction and structural validity, although there is some evidence that low rank solutions are least harmful overall. Trustworthiness seems to improve slightly as the rank is reduced, but is always poorer than FOS. In contrast, aspect level C scores become progressively worse, whilst topic level C scores follow more esoteric path, dropping until LSA-50 then rising back to approximately the same level as LSA-75PC at the lowest ranks.

Tables 3 and 4 show the F scores broken down by corpus. It is apparent the effects of increasing rank reduction vary widely between corpora. Table 3 shows the results at the topic level. In most cases, the trend is for better results to come from lower ranked solutions. Paradoxically, the two instances where LSA achieved a substantial improvement occurred after less aggressive rank reduction: T446 (LSA-90PCV) and also T307 (LSA-75PC). In contrast, the effect of rank reduction on aspect level clustering (Table 4) seems even less predictable, with low ranks being favoured by some corpora and higher ranks by others. There are only two instances where aspect level clustering is actually improved by LSA and these occurred at grossly different rank reduction levels: T387 (LSA-20) and T446 (LSA-90PCV).

| | FOS | LSA-75PC | LSA-90PCV | LSA-50 | LSA-Elb | LSA-20 | SOS | TOS |
|---|---|---|---|---|---|---|---|---|
| % Var | | 97.3 | 90.0 | 74.5 | 56.2 | 52.6 | | |
| **All** | ***.2779*** | **.2414***<br>**-13.1%** | **.2371***<br>**-14.7%** | **.2315***<br>**-16.7%** | **.2492***<br>**-10.3%** | **.2487***<br>**-10.5%** | **.2941***<br>**+5.8%** | **.2765**<br>**-0.5%** |
| T307 | *.2730* | .3013<br>+10.4% | .2504<br>-8.3% | .2433<br>-10.9% | .2531<br>-7.3% | .2531<br>-7.3% | .2765<br>+1.3% | .2434<br>-10.8% |
| T347 | *.3798* | .2956<br>-22.2% | .2925<br>-23.0% | .3135<br>-17.5% | .3687<br>-2.9% | .3585<br>-5.6% | .4021<br>+5.9% | .4123<br>+8.6% |
| T352 | *.2188* | .1742<br>-20.4% | .1696<br>-22.5% | .1686<br>-22.9% | .1936<br>-11.5% | .1891<br>-13.6% | .2440<br>+11.5% | .2298<br>+5.0% |
| T387 | *.3997* | .3133<br>-21.6% | .3153<br>-21.1% | .3366<br>-15.8% | .3570<br>-10.7% | .3593<br>-10.1% | .4345<br>+8.7% | .4102<br>+2.6% |
| T408 | *.3139* | .2726<br>-13.2% | .2664<br>-15.1% | .2796<br>-10.9% | .2929<br>-6.7% | .2973<br>-5.3% | .3319<br>+5.7% | .3177<br>+1.2% |
| T446 | *.1935* | .1819<br>-6.0% | .2155<br>+11.3% | .1508<br>-22.1% | .1434<br>-25.9% | .1483<br>-23.4% | .1878<br>-2.9% | .1634<br>-15.6% |

Table 3: Topic level F scores (overall and by corpus) and percentage improvements over FOS. Significant differences to FOS are highlighted (* p<.001).

| | FOS | LSA-75PC | LSA-90PCV | LSA-50 | LSA-Elb | LSA-20 | SOS | TOS |
|---|---|---|---|---|---|---|---|---|
| % Var | | 97.3 | 90.0 | 74.5 | 56.2 | 52.6 | | |
| **All** | ***.3887*** | **.3259***<br>**-16.2%** | **.3306***<br>**-15.0%** | **.3399***<br>**-12.6%** | **.3332***<br>**-14.3%** | **.3395***<br>**-12.7%** | **.4101***<br>**+5.5%** | **.3817**<br>**-1.8%** |
| T307 | *.5041* | .4857<br>-3.7% | .4323<br>-14.2% | .4848<br>-3.8% | .4811<br>-4.6% | .4811<br>-4.6% | .5199<br>+3.1% | .4980<br>-1.2% |
| T347 | *.2120* | .1816<br>-14.3% | .2118<br>-0.1% | .1714<br>-19.2% | .1721<br>-18.2% | .1611<br>-24.0% | .2043<br>-3.6% | .1717<br>-19.0% |
| T352 | *.3452* | .2467<br>-28.5% | .2666<br>-22.8% | .3061<br>-11.3% | .3134<br>-9.2% | .3041<br>-11.9% | .3751<br>+8.7% | .3601<br>+4.3% |
| T387 | *.4082* | .3259<br>-20.2% | .3283<br>-19.6% | .3728<br>-8.7% | .3862<br>-5.4% | .4162<br>+2.0% | .4591<br>+12.5% | .4485<br>+9.9% |
| T408 | *.3774* | .3111<br>-17.6% | .2757<br>-26.9% | .3150<br>-16.5% | .3518<br>-6.8% | .3540<br>-6.2% | .3965<br>+5.1% | .3820<br>+1.2% |
| T446 | *.4587* | .4272<br>-6.9% | .4750<br>+3.5% | .3690<br>-19.6% | .2767<br>-39.7% | .3084<br>-32.8% | .4633<br>+1.0% | .3818<br>-16.8% |

Table 4: Aspect level F scores (overall and by topic) and percentage improvements over FOS. Significant differences to FOS are highlighted (* p<.001).

In summary, the results presented here reinforce the prevailing view that there is no straightforward relationship between the singular value of a factor and its semantic salience. Whilst the top-ranked factors may often be sufficient for modelling the high-level thematic structure required for coarse partitioning tasks (Schutze and Silverstein, 1997), it seems apparent that the smallest factors play an important role in defining finer, local semantic associations (Ando, 2000; Aggarwal, 2001). The results show a dramatic drop in cluster validity occurring between full-rank and the least aggressive reduced rank condition, representing just a 3% loss of original variance. Clearly at least some of these minor factors are encoding information that is critical to the resolution of local semantic structure.

The problem still remains in determining an objective criterion that reliably identifies these optimal factors, which may be scattered across various rank positions. Whilst several criteria have been proposed (e.g. Aggarwal, 2001; Mecca et al., 2007) there still is still no generally accepted, reliable method of achieving this goal for the purpose of document similarity analysis (Janssens, 2007). One avenue that has not yet been explored, to the author's knowledge, is the possibility of adapting the concept of term discrimination value (Salton et al., 1975) as a means of selecting factors that maximise the dissimilarity between documents. However, given that LSA is a global approach to dimension reduction, it will always be essentially limited by the tendency to model features that are representative, rather than discriminatory in nature (He et al., 2004).

### *Second order similarity*

The results for SOS analysis contrast starkly to those of LSA. On average, cluster validity is improved significantly ($p < 0.001$) on all measures and at both levels of semantic association with mean gains over FOS in the region of 5 to 6% (see Table 2). At the aspect level, some 38.1% of cases show an improved F score, 48.6% of cases show no difference, whilst just 13.3% of cases show a deficit. However, the advantage is even more marked at the topic level, with some 58.1% of cases improving their F score (versus 17.5% achieving a poorer score).

It is clear from Tables 3 and 4 that the degree of improvement varies somewhat from one corpus to the next. Performance is particularly impressive for T387, where aspect and topic level cluster validity improve by 12.5% and 8.7% respectively, and for T352, where the respective gains are 8.7% and 11.5%.  The negative exceptions are T347 and T446. For T347, the topic level cluster is resolved better (+5.9%), but aspect level clusters are more poorly resolved (-3.6%). In contrast, T446 achieves only a modest improvement in aspect clustering (+1.0%) at the expense of somewhat poorer topic clustering (-2.9%).

It is interesting to note that the corpora that benefited the most from SOS analysis are those that comprise more ill-defined topical structures. Aspect overlap can be quantified by the number of aspect definitions to which each topical document relates (see Table 1), with a high average suggesting a greater tendency for aspect clusters to overlap. T352, T387 and T408 gain the biggest advantage from SOS analysis and possess a relatively high average number of aspects per topical document. In contrast, T307 and T446, which possess highly distinct aspect clusters, gain somewhat less advantage from SOS analysis.

Finally, the results of third-order similarity analysis are unimpressive in comparison, with overall performance dropping back down to just below that of FOS (not significant). Looking at the component measures, a small advantage remains in terms of trustworthiness (1-2%), but it is evident that SOS analysis is likely to be the optimal stopping point. In terms of individual corpora, reasonable gains were achieved in some cases, but these improvements were never better than those observed in the equivalent SOS model. Based on these results, higher-order similarity analyses were not attempted.

In summary, H1 is supported as SOS analysis clearly represents a reliable means of improving DSM. All corpora benefited in some way, either at the topic or aspect level. However, the corpora that obtain the greatest benefit are those composed of somewhat overlapping aspect clusters. Hence, the biggest strength of SOS analysis seems to lie in its ability to resolve the underlying structure of more complex topics.

### *Truncated SOS analysis*

Table 5 shows the test results for the truncated SOS conditions. Once again FOS is the used as the benchmark in the statistical analyses. F scores are reported for both local and global truncation schemes, whilst T and C scores are presented only for the global scheme.

As expected the global scheme shows greater robustness under increasing truncation. The advantage is particularly apparent at the topic level where the global scheme maintains a significant advantage down to K=10, whilst the usefulness of the local scheme seems to be limited to approximately K=25. The general trend, in both schemes, is one of monotonic decline in topic cluster validity with increasing truncation. The component cluster scores (global only) suggest that this trend can be attributed mainly to poorer continuity, rather than poorer trustworthiness.

A different picture emerges at the aspect level of association, with both schemes faring relatively well. The effect of local truncation is one of shallow decline with a substantial and significant advantage being maintained until K=10. Once again, however, a global threshold produces better results. In fact, global truncation can produce slightly better aspect level clustering than full vector SOS (+5.5%). An inverted-U trend is evident with optimal structure (~+6.6%) occurring in the range K=10 to K=25. Moreover, even at K=5 cluster validity remains significantly better than FOS (+6.0%). Component scores show a similar pattern to those at the topic level, with trustworthiness being more resistant to aggressive truncation than continuity and clearly responsible for the inverted-U trend seen in the F-scores.

| | | FOS | SOS-100 | SOS-50 | SOS-25 | SOS-15 | SOS-10 | SOS-5 |
|---|---|---|---|---|---|---|---|---|
| Measure | Scheme | | | | | | | |
| **F-Asp** | **Local** | **.3887** | **.4101*** | **.4080*** | **.4048*** | **.4024+** | **.4047+** | **.3844** |
| | **%** | | **+5.5** | **+5.0** | **+4.1** | **+3.5** | **+4.1** | **-1.1** |
| | **Global** | **.3887** | **.4101*** | **.4100*** | **.4145*** | **.4146*** | **.4138*** | **.4120*** |
| | **%** | | **+5.5** | **+5.5** | **+6.6** | **+6.7** | **+6.5** | **+6.0** |
| T-Asp | Global | .3950 | .4216* | .4194* | .4266* | .4252* | .4201* | .4216* |
| | % | | +6.7 | +6.7 | +8.0 | +7.6 | +6.4 | +6.7 |
| C-Asp | Global | .5482 | .5703* | .5751* | .5724* | .5653* | .5646* | .5604 |
| | % | | +4.0 | +4.9 | +4.4 | +3.1 | +3.0 | +2.2 |
| **F-Top** | **Local** | **.2779** | **.2941*** | **.2891*** | **.2816+** | **.2766** | **.2736** | **.2594*** |
| | **%** | | **+5.8** | **+4.0** | **+1.3** | **-0.5** | **-1.6** | **-6.7** |
| | **Global** | **.2779** | **.2941*** | **.2934*** | **.2896*** | **.2848*** | **.2815+** | **.2822** |
| | **%** | | **+5.8** | **+5.6** | **+4.2** | **+2.5** | **+1.3** | **+1.5** |
| T-Top | Global | .5137 | .5448* | .5416* | .5416* | .5371* | .5333+ | .5308+ |
| | % | | +6.1 | +5.4 | +5.4 | +4.6 | +3.8 | +3.3 |
| C-Top | Global | .2076 | .2197* | .2200* | .2151* | .2120 | .2081 | .2116 |
| | % | | +5.8 | +6.0 | +3.6 | +2.1 | +0.2 | +1.9 |

Table 5: T, C and F scores (aspect then topic level) and percentage improvements of SOS and truncated SOS (K=50 to K=5 over FOS. Significant differences to FOS are highlighted (* p<.001; + p<.05).

Hence H2 is supported as a global threshold is clearly the more efficacious scheme, at all mean truncation levels. The main penalty of global truncation on SOS analysis seems to be a loss of cluster continuity. In contrast, cluster trustworthiness holds up remarkably well, even under quite extreme truncation. This is particularly evident at the aspect level of association. In short, the general tendency is for the higher-level themes (i.e. topics) to become fragmented into more focused albeit isolated clusters. This is to be expected, given that the truncation process is one where the effects of the majority of weaker FOS coefficients give way to those of the minority of more salient inter-document associations.

If the goal is to improve run-time whilst retaining the general characteristics of a full-vector SOS analysis, then K=25 seems to offer the optimal trade-off. Generally, at this level, a slight loss of topic level cluster quality is offset by gains at the aspect level of association. Clearly, further testing is required to determine whether this rule can be generalized to other kind of document corpora. However, in the Spatialisation section below, we demonstrate that the K=25 heuristic also produces excellent results for the WOS3T dataset, a much larger corpus composed of academic abstracts, rather than news articles.

In contrast, if the objective is geared more towards resolving topics at finer levels of abstraction, then it seems feasible and potentially advantageous (in terms of speed and quality) to apply a substantially more aggressive truncation threshold. It must be noted, however, that whilst the results for K<=10 look impressive, such aggressive truncation does carry a high risk of disconnection within the similarity graph. All of the K=5 models created for this study comprised at least one disconnected document case (where similarity to all other documents was zero). For spatialisation and some clustering applications, it is essential for the DSM to be a fully connected graph. The minimum K for these corpora ranged from 10 to 25, with 10 being the mode (4 out of 6 corpora). T352 required at least K=15, whilst T408 required a minimum of K=25. Unfortunately, there seems to be no simple way of predicting minimum truncation level. There is no correlation between corpus size and minimum K; connected graphs for most of the smaller corpora can be resolved down to K=10, yet the smallest was the only one that required K=25. In fact none of the known properties (Table 1) of these corpora seemed to reliably predict minimum K. In reality, minimum K is likely to depend upon local anomalies (e.g. unusually distinct documents) more than global characteristics. Typically when minimum K is reached it is just a handful (often just one or two) of documents became disconnected from the main, otherwise intact graph. It is likely that the problem can be alleviated, to some extent, by placing a lower-bound on the number of non-zero elements that must be retained per document column i.e. to retain certain elements even if their values fall below the threshold.

In summary, the results show that a global truncation scheme produces better results than a location scheme. It is generally possible to achieve comparable results to full-vector SOS analysis down to K=25, indicating that speed-ups of at least four times are realistically achievable. Moreover, in cases where global continuity is less important, speed-ups of ten or more times are feasible. Whilst the computational complexity associated with determining the exact global threshold proves an excessive overhead (actually slowing down the SOS computation), in the next sub-section we demonstrate two efficient approaches to establishing the global threshold that enable truncated SOS to achieve speed-ups inline with theoretical expectation.

### *Run-time analysis*

The theoretical speed-up of the truncation method is inversely proportional to K, which is the average percentage of non-zero elements retained in each document vector. However, it was clear during cluster validity testing that the overhead of computing the global threshold prior to truncation tends to outweigh the resulting gains during the SOS computation step. As such, two faster methods of establishing the threshold, *global mean* and *hybrid* are proposed and tested later in this section.

Table 6 shows the SOS relative computation times of the various methods discussed across 12 corpora. These include the six FT corpora as used in the earlier analyses plus six additional datasets, the five *MT* corpora (N=100 to N=500) and *WOS3T* (N=2235).

| ***N*** | ***100*** | ***122*** | ***127*** | ***137*** | ***162*** | ***200*** | ***218*** | ***300*** | ***400*** | ***500*** | ***588*** | ***2235*** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Full** | **.005** | **.010** | **.012** | **.013** | **.024** | **.041** | **.055** | **.155** | **.478** | **1.288** | **2.882** | **312.79** |
| G=25 | .10 | .09 | .09 | .08 | .07 | .06 | .05 | .06 | .07 | .01 | .06 | .01 |
| $K_{GM}$ | *24.56* | *28.53* | *24.00* | *24.93* | *27.27* | *27.67* | *31.18* | *29.21* | *30.13* | *30.35* | *31.97* | *41.76* |
| GM | 3.00 | 2.94 | 3.90 | 3.37 | 3.42 | 3.62 | 3.17 | 3.67 | 4.26 | 5.34 | 4.79 | 2.64 |
| $H_{100}$=25 | .67 | .86 | .91 | .83 | .89 | .87 | .96 | .94 | 1.30 | 1.79 | 2.19 | 2.53 |
| $H_{20}$=50 | 1.00 | 1.26 | 1.41 | 1.15 | 1.30 | 1.24 | 1.29 | 1.43 | 1.64 | 2.07 | 2.09 | 1.96 |
| $H_{20}$=25 | 1.78 | 2.37 | 2.66 | 2.10 | 2.43 | 2.29 | 2.42 | 2.68 | 3.31 | 4.37 | 4.54 | 4.13 |
| $H_{20}$=15 | 2.53 | 3.46 | 4.18 | 3.05 | 3.75 | 3.58 | 3.66 | 4.18 | 5.17 | 6.87 | 7.61 | 7.33 |
| $H_{20}$=5 | 5.33 | 7.46 | 9.00 | 8.00 | 9.44 | 9.20 | 11.68 | 11.55 | 14.85 | 20.61 | 24.28 | 27.27 |

Table 6: Relative run-time of truncated SOS across twelve corpora, presented in order of size from N=100 to N=2235. *Full* row shows actual time in seconds required by full-vector SOS analysis. $K_{GM}$ row shows

the mean number of non-zero elements after global mean truncation. All other rows show speed-up achieved by different methods relative to *Full*.

The row labelled *Full* shows the absolute time (in seconds) required to compute a full vector SOS analysis for each dataset. Even though the test PC is a fairly modest platform by current standards, computation time is clearly not an issue for corpora up to ~500 documents in size. Inline with the cubic complexity, the MT200 matrix takes approximately eight times longer to compute than MT100. However, the run-time growth-rate clearly escalates once N passes beyond 500 documents. For instance, the time required to compute the SOS matrix for T446 (N=588) is over twice that required for MT500. We would expect the increase to be closer to 1.6 times based on cubic complexity. Moreover, for the WOS3T corpus, extrapolation from the times for the small datasets leads us to expect that SOS analysis should take somewhere between 115 – 158 seconds. In fact, the computation took 313 seconds, over twice as long as expected.

This observation is probably a result of an increased requirement for data swapping between RAM and the faster, but limited cache memory (L2 = 512KB, L3 = 2MB). Janssen (2007) doubted the feasibility of SOS analysis for corpora of more than a few thousand documents, apparently on the grounds of cubic complexity alone. It seems likely that the finite capacity of CPU caches may pose an additional limitation upon the procedure. However, it should be possible to bring computation times down significantly by using more efficient data structures. In our implementation, whilst indexes were used to provide fast look-up of non-zero elements, the values themselves were actually held in a standard two-dimensional array, which is a somewhat inefficient way of storing a sparse matrix. Storing the data in a sparse vector format (e.g. an array of lists) would reduce the memory overhead considerably.

The cost of computing the exact global threshold is clearly prohibitive. The second row labelled *G = 25* shows the relative speed of truncated SOS (K=25) over full SOS analysis when using this method. Relative run-times are ten times higher when N=100, and increase steadily with corpus size. It is clear that faster methods of computing the threshold are required.

The first alternative method is simply to use the mean of all values in the FOS matrix of establishing the threshold. The *global mean* method has almost no cost as it simply involves summing values as the FOS matrix is computed or retrieved from file, followed by a single division. Fortunately, this criterion results in a surprisingly consistent K of around to 24% to 32% (see row $K_{GM}$ in Table 6) for all of the Financial Times news article corpora, which is close to the optimal truncation level identified previously. The WOS3T corpus did yield a somewhat higher average of 42%. This may reflect the fact that academic articles are written using a more controlled vocabulary than news articles and also that other factors such as the document (abstract) length will be less variable. The implication, however, is that the global mean heuristic may be less useful in the case of more homogeneous corpora

Relative speed of computation is shown in the row labelled *GM*. The speed-ups are generally inline with expectations. It is apparent that the relative gains increase with corpus size as the additional costs associated with computing the threshold and pruning the matrix become progressively more outweighed by decreased workload at the SOS analysis step. It can be noted that for the larger datasets the speed-ups seem slightly greater than expected. This is actually a consequence of the fact that, using a global threshold, the distribution of per document truncation rates tends to be positively skewed, so whilst the mean truncation rate is equal to K%, considerably more than 50% of vectors will be truncated more aggressively than this.

The second alternative method of establishing the global threshold is a *hybrid* of the local and global method. In this approach, the local thresholds for a sample of FOS vectors are computed and the mean of these values is taken as an estimate of the global threshold. The row labelled $H_{100}$=25 (Table 6) shows the times for the hybrid method using a sample of 100% (mean of all local thresholds). The resulting times are still slower than full SOS up to N=300, although it can also be seen that the speed-up gradually converges towards the expectation as corpus size increases. Much better results are achieved when the threshold is estimated from a smaller sample of vectors. The remaining rows of the table, show the results using a sample rate of 20%, (i.e. $H_{20}$ = K). This setting is five times faster yet still offers reasonably accurate control over K, normally within five percent of the target value. Moreover, computation times meet or exceed the theoretical speed-ups for datasets of 500 documents or more, whilst useful speed-ups are still achieved on the smaller sets. Reducing the sampling rate further may improve the speed for smaller corpora, albeit at the cost of less precise control over K.

### *Visual comparison*

The preceding analyses have demonstrated significant gains in cluster validity (F, T and C). However, statistical analysis does not fully illustrate the extent of structural change imposed by SOS. Whilst a mean gain of 5-6% sounds relatively modest, these figures tend to obscure the impact of a minority of large improvements in the topology of the similarity graph. In this sub-section, DSMs are visualised using spatialisation (distance-similarity visualization) created with the SPSS PROXSCAL procedure using an ordinal scaling criterion. Figures 3 and 4 show spatialisations created from the FOS, SOS and SOS-GM models of the T352 and T387 corpora. In Figure 3, all topical documents are marked-up in black. In Figure 2, colour is used to differentiate selected aspect clusters within these corpora.

Figure 3 shows that whilst some outliers remain after SOS analysis, the majority of relevant documents converge to form a kind of 'bulls-eye' feature. Moreover, SOS-GM produces a very similar configuration. This bulls-eye effect is highly desirable both from a visualization perspective, as the user's attention will be naturally drawn to the most densely populated regions of the space (Rorvig and Fitzpatrick, 1998; Rorvig, 1998; Hornbaek and Frokjaer, 1999; Montello et al., 2003). The density differential also provides the opportunity to implement effective, automatic relevance feedback systems based on the corpus centroid or medoid (Rorvig, 1998). Rorvig (op cit.) found that query surrogates generated by using known relevant document vectors from the densest part of the similarity space can increase the recall rate by up to 46%.

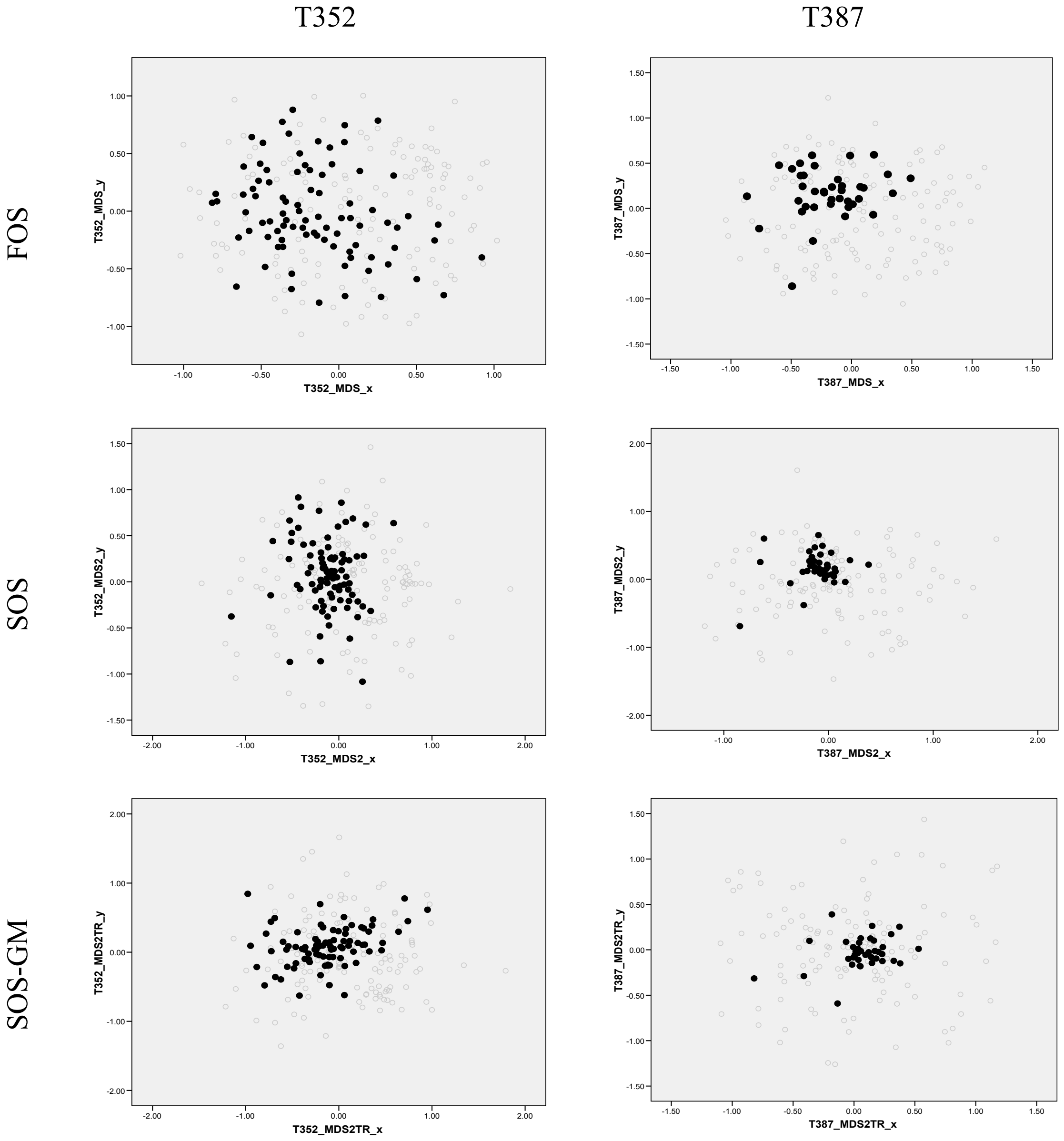


Figure 3: MDS spatialisations for T352 and T387 with topical documents highlighted.

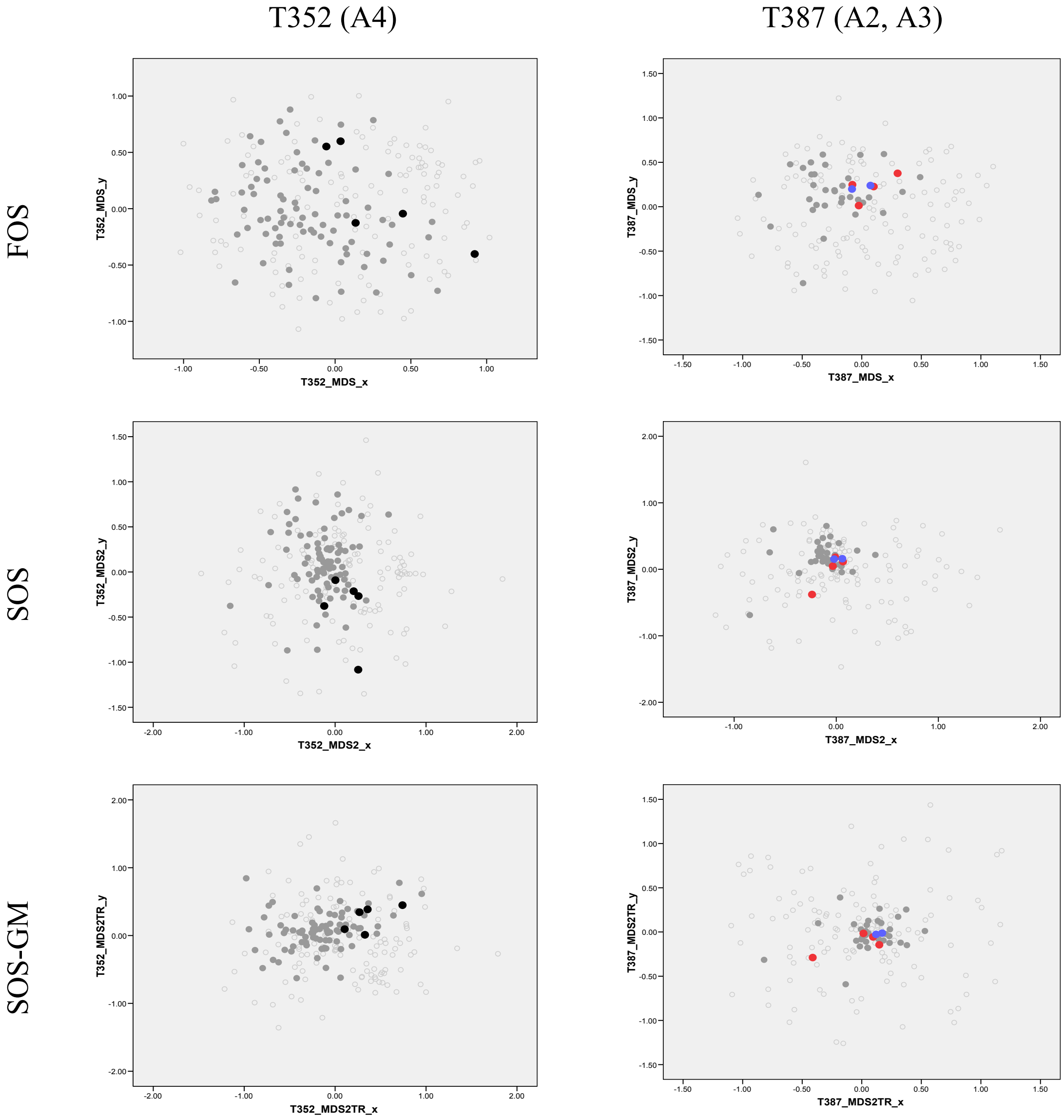


Figure 4: MDS spatialisations for T352 and T387 corpora with topical documents shown in solid mid-grey and aspect relevant documents shown in black (T352, A4), blue (T387, A2/A3) and red (T387, A2).

Figure 4 illustrates the spatial-semantic effect of SOS at the finer, aspect level of association. The first case (left column) relates to a specific aspect of the T352 corpus. The topic description of T352 is as follows:

*Impacts of the Chunnel anticipated or actual on the British economy and/or the life style of the British*

As mentioned previously, the aspect definitions of T352 are relatively ill-defined and there is a high degree of overlap between the sub-clusters. SOS and SOS-GM achieved large overall improvements in aspect level F scores of 8.7% and 14.0% respectively. The news archive used in the TREC competition covers a period that spans the final stages of the Chunnel's construction, its completion and opening, and the immediate aftermath of its opening during 1994. Hence, a wide variety of issues relevant to the topic description were reported by the press during this period. One of the key impacts of the Chunnel was the threat to the business of the ferry operators, P&O and Stena-Sealink, who responded by requesting permission to merge their Dover-Calais operations.

According to the TREC relevance data, five documents in this corpus are relevant to this aspect. Documents #67 and #115 are closely related as they were both published in 1992 and discuss the operators' pre-emptive request to merge, which was rejected at that point due to delays that had occurred in the Chunnel opening. Both FOS and SOS effectively identify this strong association which can be seen in all three spatialisations as the closest pair of black nodes.

The three remaining aspect members are located some distance away from within the FOS spatialisation. The next closest pair, #1 and #163 are articles published some time later, around the time of opening in 1994 and discuss the likelihood of price-wars occurring once the Shuttle (car-train) goes into service. Whilst FOS fails to resolve the semantic association between these two document pairs, they are much more proximal in the SOS and SOS-GM spatialisations. Moreover, these four aspect members all converge neatly on the bottom edge of the dense main cluster, making them more readily discoverable by a browsing user.

One final lone node, #106, remains somewhat isolated even after SOS analysis. This document is clearly associated with the others, referring to another attempt made by the ferry operators in 1993 to gain permission to merge, but is particularly short, with only 38 words making up both the title and main body combined. The SOS-GM solution appears to resolve the aspectual association slightly better than SOS. This may be an artefact of the MDS process however; both SOS and SOS-GM manage to increase aspect level trustworthiness, over FOS, to the same extent (from 0.2 to 0.4). Strangely, aspect level C scores are equal to one (the maximum score) under all similarity conditions so, theoretically, #106 should not be an outlier. Again, this may reflect a limitation of the MDS process, which attempts to achieve a globally optimal solution. Being a short document means that the FOS (and SOS) profile of #106 is relatively more sparse than most and contains lower cosine coefficients. It is likely that a non-linear projection scheme (e.g. ISOMAP: Tenenbaum et al., 2000) that focuses more on preserving local associations would result in a better spatial location for such a document (see Cribbin, 2010).

The right-hand column of Figure 4 shows a second case of aspect clustering, this time in the T387 corpus. This is quite a different example, comprising two, overlapping clusters. The topic description for T387 is as follows:

*Identify documents that discuss effective and safe ways to permanently handle long-lived radioactive wastes.*

Aspects two (A2) and three (A3) relate to the encasement of such waste in concrete and steel respectively. In this corpus, the overlap occurs because both members of A3 are also members of A2, discussing both concrete and steel methods. These two 'dual-aspect' document nodes are highlighted in blue. The four other members of A2 are discuss only steel methods of encasement and are highlighted in red. As before, solid grey nodes indicate other documents relevant to this topic.

Given this aspectual structure, we would expect all six documents to group together, with the blue overlapping nodes forming a sub-cluster within. Whilst there is some evidence of clustering in the FOS spatialisation, the desired formation is more evident in the SOS solutions, with five out of the six documents clearly converging within the dense region defined by the general topic. The relationship between the two A3 documents is well represented in both SOS solutions, although this aspect seems to be distinguished slightly more clearly by SOS-GM.

In all of the T387 spatialisations there is one clear outlying document: #80. This document is pushed further from the main topic cluster in both the SOS spatialisations. However, this seems to make some sense as it is quite distinct from most other topic documents. Firstly, whilst most (30 out of 38) of the relevant documents discuss underground disposal methods, #80 focuses on the disposal of waste above ground. Secondly, most of the topical documents refer to nuclear projects in the UK and Europe, whilst this one focuses on a Russian project. It is also notable that #80 achieved a zero trustworthiness score at both the topic and aspects level of association in all three models. Its continuity score was also zero at the aspect level and very low at the topic level, although both SOS transformations did manage to slightly increase the document's topic continuity score (from 0.105 to 0.132).

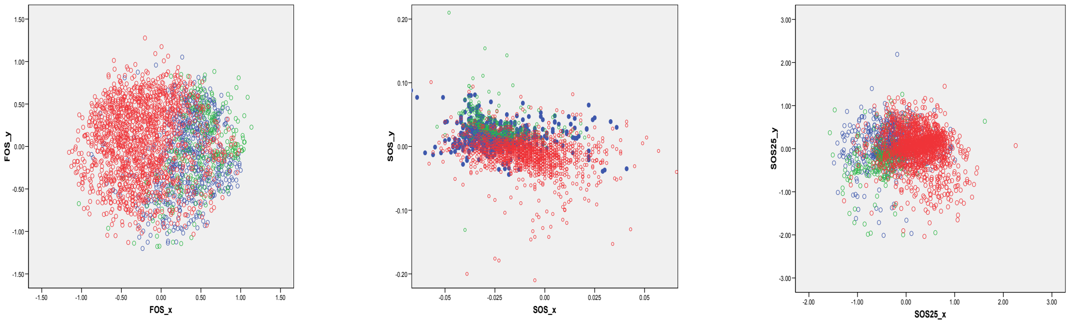
An important question lies with respect to the scalability of the truncation method. The next example, WOS3T, is a large corpus of academic abstracts (N=2235) retrieved from the Web of Science (Figure 5). This is not a well-defined topical subset like the previous examples, but is composed of all the articles published by three IEEE Transactions journals (Education, Computer, Nuclear Science) over a five year period (2005-9). This represents a somewhat different and more challenging similarity modelling task. Being IEEE journals, all three subsets are related by their engineering/technology focus. Hence one would expect some significant degree of semantic overlap between the subsets (e.g. Both Nuclear Science and Computer articles might discuss applications of simulation modelling techniques). On the other hand, each journal has a different focus audience and will therefore tend to address quite distinct problems.

Nuclear science is the dominant subset (1394 documents) and perhaps the most distinct. In the FOS solution the majority of these documents (shown as red nodes) form a distinct cluster that dominates the left side of the visualization. The right-hand side of this cluster then merges into the space containing the remaining two subsets. It is possible to make out a definite, almost straight line demarcating main Nuclear Science cluster from the Education (green nodes) and Computers (blue nodes) clusters. It seems that Education articles tend towards the top right whilst Computers articles tend more towards the bottom right. However, there is no clear partition between these two topics.

The SOS spatialisation is quite different. This is a particularly extreme example of the SOS 'bulls-eye' effect, whereby whilst the majority of documents converge to the centre of the space, a minority of more distinct documents are dispersed widely. Such was the extent of this centrifugal effect that, when preparing this figure, it was necessary zoom deeply into the original PROXSCAL solution in order to present a sufficient level of detail on the main cluster. This is a clear reminder of how SOS analysis not only causes similar documents to converge but dissimilar ones to diverge, which is consistent with the known effect of higher-order similarity analysis (Chen, 2002). As before, the majority of Nuclear Science articles converge into a single neat cluster. A distinct 'orphan' sub-set is also now evident beneath this main cluster. Perhaps more interesting is the change in structure of the two smaller subsets. The Education articles are now more effectively partitioned within the space, with the majority residing in a distinct cluster towards the top-left of the main Nuclear Science cluster. There is no clear cluster defining the Computers articles, however. These articles are still distributed widely across the top edge of the main Nuclear Science cluster.

The truncated (SOS-25) solution instantly seems more satisfying. For the first time, both the Computers and Education articles both have their own, distinct regions of the space. Whilst the majority of articles in Nuclear Science subset still form a distinct and dense cluster, the outlying cluster of 'orphan' articles, seen in the SOS solution, is still clearly evident. Moreover, although some outliers remain, the extreme cases occurring in the full-vector SOS solution are not evident in this case.

In summary, the WOS3T corpus is a compelling case where truncated SOS produces an arguably more plausible spatial-semantic configuration than full-vector SOS. It is not clear exactly why this should be, although it should be noted that the term-document space of this corpus, comprised of academic articles, is significantly more homogeneous than those of the FT corpora. Mean FOS is 0.13 (versus ~0.05 or lower for FT) and similarities are more normally distributed, with some 42% of elements falling above the mean (see Table 6). However, recalling how truncation was particularly effective at resolving aspectual clusters in the FT corpora, this may simply be good example of truncation magnifying finer relationships at the expense of more general ones. This ability to distil large, homogeneous corpora suggests that truncated SOS analysis can make a valuable addition to the text-miner's toolbox.



FOS SOS SOS-25

Figure 5: Spatialisations of models derived from the WOS3T corpus. Red nodes denote *Nuclear Science* articles, blue nodes *Computers* and green nodes *Education*.

## Final discussion and conclusions

This paper has examined the effect of second-order similarity analysis on the quality of clustering in document similarity models computed using a bag of words vector space model. Differences with respect to both first-order and LSA derived similarity models are reported. This study extends previous work (Ahlgren and Colliander, 2009; Janssen, 2007) in the area in three main ways. Firstly, rather than examining simple post-cluster classification accuracy, document similarity models were evaluated using a variant of the nearest-neighbours test incorporating measures of both trustworthiness and continuity at two levels of semantic association. Secondly, corpora of varying size and structural characteristics were examined, rather just a single corpus. Thirdly, a solution to extant concerns over scalability (e.g. Janssen, 2007) was proposed in the form of vector truncation, a method that has been applied elsewhere as a means of improving run-time performance of document clustering (e.g. Schütze and Silverstein, 1997).

Contrary to previous studies, the results presented here suggest that LSA actually harms rather than improves the quality of clustering, whilst SOS analysis produces consistent structural improvements. This disagreement is attributed to the methodology used to evaluate cluster quality. Both Janssen (2007) and Ahlgren and Colliander (2009) evaluated these similarity schemes based on the extent to which documents are assigned to the 'correct' category by some clustering algorithm. In contrast, this study applied a variation of the nearest-neighbour test (Voorhees, 1985; Cribbin, 2010) which provides an ordinal measure of the extent to which classified documents are located near to related documents within the DSM itself. It may be that LSA has been afforded an unfair advantage in earlier studies, as discrete clustering algorithms are likely to favour dimension reduction approaches, like LSA, that emphasise representative features over discriminative ones (He et al., 2004).

Although it is generally acknowledged that clustering works best using low rank LSA solutions (e.g. Janssen, 2007; Schütze and Silverstein, 1997), no clear correlation was found between the number of factors retained and the quality of the DSM. Whilst the lower rank conditions did result in slightly better trustworthiness, continuity was not improved and generally performance was inferior to that of the simple bag of words method. A large disparity in performance between the highest rank (representing 97% of original variance on average) and the full rank condition suggested that useful structural information resides within even the smallest factors. Indeed, a consequence of the global objective of SVD, which attempts to capture as much variance as possible per factor, may be that the higher-ranked factors, by definition, tend to over-emphasise thematic relationships (He et al, 2004), obscuring important yet subtle topical distinctions in the process (Ando, 2000; Aggarwal, 2001).

In contrast H1, which postulated that second-order similarity analysis can reliably detect latent semantic structure that is invisible to FOS analysis, is supported. Inline with previous studies (Janssens, 2007; Ahlgren and Colliander, 2009) the quantitative analyses showed that SOS analysis can improve cluster quality at both high and low levels of semantic association. SOS is particularly adept at resolving clusters within complex topics i.e. where distinctions between aspects of the topic are less clearly defined. MDS spatialisations illustrate the full impact of the SOS transformation, revealing better spatial partitioning between the expected clusters. Of particular interest is the tendency for documents relating to the dominant theme of a corpus to converge into a dense 'bull's-eye' type structure near to the centre of the visualization. This is not only desirable from the visualization perspective but may also have useful applications as a pseudo-relevance feedback device in IR systems (see Rorvig, 1998). The potential of the latter application forms an interesting subject for future research.

Although it may be tempting to assume that further iterations of similarity analysis might lead to even better similarities (see for e.g. Chen, 2002), this was not borne out by the evaluation of third-order similarity models (i.e. computing similarities in terms of second-order similarity profiles). The results showed that the additional iteration actually harms rather than improves cluster structure.

One of the key problems of SOS analysis has been its scalability (Janssen, 2007). Whilst the cubic complexity remains, it has been demonstrated how, by truncating the FOS vectors prior to SOS analysis, it possible to achieve constant improvements in run-time of at least four times, without incurring significant structural penalties. The main penalty associated with increasing truncation rate is one of reduced continuity, particularly in terms of clustering more general themes. However, as predicted by H2, this effect was considerably less marked when vectors were truncated according to a global, rather than locally determined threshold. Moreover, global truncation often resulted in improvements to the general trustworthiness of the lower level semantic structures. In short, the general effect of truncation appears to be to disperse and distil the main thematic clusters into their component aspects.

Whilst the cost of determining the exact global threshold is excessive, a good estimate can be obtained by taking the mean of a small sample of local thresholds. Using this hybrid sampling approach makes it possible to achieve the theoretical constant improvements in run-time. A safe heuristic to apply when exploring uncharted corpora is a threshold of 25% (i.e. to zero the bottom 75% of similarities). This typically results in structures that are as good if not better than those of the full-vector method, whilst usefully reducing run-time by a factor of four.

Scope clearly exists for more aggressive truncation (<10%) which, in addition to greater speed-ups, has been shown in many cases to lead to improved resolution of low level semantic features. However, there is also a risk of the similarity graph becoming disconnected when the threshold drops substantially below 25%. Future development of the method should identify a means of attenuating this risk, for example by setting a lower-bound of non-zero elements per vector, without sacrificing the performance advantages observed here.

Finally, there are many opportunities to improve the scalability of SOS analysis still further. For instance, the use of sparse-vector data structures would allow better use of CPU caches which would improve performance on larger corpora. Also, the nature of algorithms like LSA and SOS analysis make them prime candidates for parallelisation, using GPU hardware. Cavanagh et al. (2009) recently demonstrated a GPU version of LSA that delivered speed-ups in the order of five to six times. Libraries for basic linear algebra (BLAS) are already available for compute platforms such as Nvidia's CUDA[2], so porting (truncated) SOS analysis is a logical and relatively simple next step. As GPU hardware and programming methods evolve, it will become feasible to apply SOS analysis to significantly larger corpora than those analysed to date.

[2] 1. http://developer.nvidia.com

## References

Aggarwal, C. C. (2001). *On the effects of dimensionality reduction on high dimensional similarity search.* Paper presented at the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, Santa Barbara, California, United States.

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics, 3*(1), 49-63.

Ando, R. K. (2000). *Latent Semantic Space: Iterative Scaling Improves Inter-document Similarity Measurement.* Paper presented at the SIGIR 2000.

Cavanagh, J. M., Potok, T. E., & Cui, X. (2009). *Parallel latent semantic analysis using a graphics processing unit.* Paper presented at the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers, Montreal, Québec.

Chen, C. (1998). Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers, 10*(2), 107-128.

Chen, C. (2002). Generalized association plots: Information visualization via iteratively generated correlation matrices. *Statistica Sinica, 12*, 7-29.

Chen, C., & Czerwinski, M. (1998). *From latent semantics to spatial hypermedia: An integrated approach.* Paper presented at the 9th ACM Conference on Hypertext (Hypertext '98), Pittsburgh, USA.

Cribbin, T. (2010). Visualising the structure of document search results: a comparison of graph theoretic approaches. *Information Visualization, 9*(2), 83-97.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964-971.

He, X., Cai, D., Liu, H., & Ma, W.-Y. (2004). *Locality preserving indexing for document representation.* Paper presented at the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom.

Hornbæk, K., & Frokjær, E. (1999). *Do Thematic Maps Improve Information Retrieval.* Paper presented at the IFIP TC.13 International Conference on Human-Computer Interaction (INTERACT '99).

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics.* Unpublished Doctoral Dissertation, Katholieke Universiteit, Leuven.

Kontostathisa, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing and Management, 42*(1), 56-73.

Landauer, T., & Dumais, S. (1994). *Latent Semantic Analysis and the measurement of knowledge.* Paper presented at the Proceedings of the First Educational Testing Service Conference on Applications of Natural Language Processing in Assessment and Education.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. E. (1997). *How well can passage* *meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans.* Paper presented at the Proceedings of the 19th annual meeting of the Cognitive Science Society, Mawhwah, NJ.

Larsen, B., & Aone, C. (1999). *Fast and effective text mining using linear-time document clustering.* Paper presented at the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego.

Lee, M., Pincombe, B. M., & Welsh, M. B. (2005, July 21-23). *An empirical evaluation of models of text document similarity.* Paper presented at the 27th Annual Conference of the Cognitive Science Society, Stresa, Italy.

Madsen, R. E., Sigurdsson, S., & Hansen, L. K. (2004). *Enhanced Context Recognition by Sensitivity Pruned Vocabularies.* Paper presented at the 17th International Conference on Pattern Recognition (ICPR 2004).

Martín-Merino, M., & Muñoz, A. (2004). *A New MDS Algorithm for Textual Data Analysis.* In Neural Information Processing (pp. 860-867). Berlin / Heidelberg: Springer.

McCain, K. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433-443.

Mecca, G., Raunicha, S., & Pappalardo, A. (2007). A new algorithm for clustering search results. *Data & Knowledge Engineering, 62*(3), 504-522.

Montello, D. R., Fabrikant, S., Ruocco, M., & Middleton, R. S. (2003). Testing the First Law of Cognitive Geography on Point-Display Spatializations. *In Spatial Information Theory: Foundations of Geographic Information Science* (Vol. Lecture Notes in Computer Science 2825, pp. 316-331). Berlin: Springer-Verlag.

Muresan, G., & Harper, D. (2004). Topic modeling for mediated access to very large document collections. *Journal for the American Society for Information Science and Technology, 55*(10), 892-910.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846-850.

Rorvig, M. (1998). Scaled structure in visualized TREC data and query feedback. *Information Processing & Management, 34*(2-3), 151-160.

Rorvig, M., & Fitzpatrick, S. (1998). Visualization and scaling of TREC topic document sets. *Information Processing & Management, 34*(2-3), 135-149.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill Inc.

Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM, 18*(11), 613-620.

Schütze, H., & Silverstein, C. (1997). *Projections for efficient document clustering.* Paper presented at the 20th annual international ACM SIGIR conference on Research and development in information retrieval, Philadelphia, Pennsylvania, United States.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science, 290*(5500), 2319-2323.

van Rijsbergen, C. (1979). *Information Retrieval*. London: Butterworths.

Venna, J., & Kaski, S. (2006). Local multidimensional scaling. *Neural Networks, 19*(6-7), 889-899.

Voorhees, E. (1985). *The cluster hypothesis revisited.* Paper presented at the 8th annual international ACM SIGIR conference on Research and development in information retrieval, Montreal, Quebec, Canada.

Voorhees, E., & Harman, D. (1997). *Overview of the Sixth text REtrieval Conference (TREC-6).* Paper presented at the Sixth Text REtrieval Conference (TREC-6), Gaithersburg, Maryland.

Voorhees, E., & Harman, D. (1998). *Overview of the Seventh Text REtrieval Conference (TREC-7).* Paper presented at the Seventh Text REtrieval Conference (TREC 7), Gaithersburg, Maryland.

Voorhees, E., & Harman, D. (1999). *Overview of the Eighth Text REtrieval Conference (TREC-8)*. Paper presented at the Eighth Text REtrieval Conference (TREC-8), Gaithersburg, Maryland.