Visualising the structure of document search results: a comparison of graph theoretic approaches

Timothy Cribbin

Department of Information Systems and Computing, Brunel University, Uxbridge, UK.

UB8 3PH.

timothy.cribbin@brunel.ac.uk

Abstract: Previous work has shown that distance-similarity visualisation or 'spatialisation' can provide a potentially useful context in which to browse the results of a query search, enabling the user to adopt a simple local foraging or 'cluster growing' strategy to navigate through the retrieved document set. However, faithfully mapping feature-space models to visual space can be problematic due to their inherent high dimensionality and non-linearity. Conventional linear approaches to dimension reduction tend to fail at this kind of task, sacrificing local structural in order to preserve a globally optimal mapping. In this paper the clustering performance of a recently proposed algorithm called isometric feature mapping (Isomap), which deals with non-linearity by transforming dissimilarities into geodesic distances, is compared to that of non-metric multidimensional scaling (MDS). Various graph pruning methods, for geodesic distance estimation, are also compared. Results show that Isomap is significantly better at preserving local structural detail than MDS, suggesting it is better suited to cluster growing and other semantic navigation tasks. Moreover, it is shown that applying a minimum-cost graph pruning criterion can provide a parameter-free alternative to the traditional K-neighbour method, resulting in spatial clustering that is equivalent to or better than that achieved using an optimal-K criterion.

This is a post-peer-review, pre-copyedit version of an article published in *Information Visualization*. The definitive publisher-authenticated version of **Cribbin, T. (2010). Visualising the structure of document search results: a comparison of graph theoretic approaches. *Information Visualization, 9*(2), 83-97** is available online at: http://dx.doi.org/10.1057/ivs.2009.3

Keywords: information retrieval; document visualisation; multidimensional scaling; isometric feature mapping; minimum spanning tree; pathfinder associative network

1. Introduction

Spatialisation is an approach that seeks to visualise the salient structure of a high-dimensional feature space as a map wherein inter-document distances correlate inversely with their similarity in input space. The distance-similarity metaphor is desirable because it exploits our innate tendency to form a correspondence between perceived spatial proximity of objects and their conceptual similarity (Tobler, 1970; Montello et al., 2003).

This metaphor presents a particularly attractive solution to the problem of navigating or exploring a set of unfamiliar documents, for example a set of search results. The cluster hypothesis of information retrieval states that documents that are proximally located in feature space will tend to be relevant to the same queries (van Rijsbergen, 1979). This leads to the logical hypothesis that by visualising the feature space in two or three dimensions using MDS or some other method of dimension reduction, documents that are relevant to the query should self-organise into a single, discernable cluster. There have been several examples that demonstrate the feasibility of this concept using both discrete clustering (Hearst and Pederson, 1996; Wu et al., 2001) and MDS (Allan et al., 2001; Cribbin and Chen, 2001).

If mainstream search interfaces are to successfully exploit spatialisation methods, it is important to define algorithms that can consistently achieve visual structures that possess this property with little or no supervision. However, previous work has demonstrated that this is not a simple goal. Feature spaces generated using automatic text analyses tend to be high-dimensional and sparse in structure, which means that are large proportion of document pairs have little or no measurable similarity. Proximity distributions that are severely skewed in this way do not lend themselves well to automation using traditional

This is a post-peer-review, pre-copyedit version of an article published in *Information Visualization*. The definitive publisher-authenticated version of **Cribbin, T. (2010). Visualising the structure of document search results: a comparison of graph theoretic approaches. *Information Visualization, 9*(2), 83-97** is available online at: http://dx.doi.org/10.1057/ivs.2009.3

linear methods (Martin-Merino and Munoz, 2004). In this paper, we explore the efficacy of a promising non-linear dimension reduction method called Isomap (Tenenbaum et al., 2000). Isomap has been shown, in other domains, to render good cluster separation when the data is inherently non-linear or 'noisy' in structure. This is achieved by first transforming the inter-document proximities into geodesic distances before inputting the new matrix into a conventional MDS algorithm. Although simple to understand and efficient to implement, to date there has been surprisingly little evaluation of the algorithm as a document spatialisation method, despite calls for its evaluation (Skupin and Fabrikant, 2003). Existing studies have only studied single document sets using qualitative inspection (Navarro and Lee, 2001) and user-search performance (Butavicious and Lee, 2007) as the measures of structural fidelity. No known study has attempted to make an objective comparison of cluster integrity between non-linear and linear spatialisation solutions across a range of different document sets.

The experiments reported here evaluate spatialisation performance across 16 distinct document sets. An adaptation of Voorhees' (1985) nearest neighbours test is used to quantitatively evaluate topic clustering in visualisations created using MDS and a number of variations of the Isomap method. Results show that Isomap returns significant improvements in cluster separation, but that observation of this advantage depends on the homogeneity of the expected clusters (i.e. the relevance criterion used). A second contribution of this paper is to compare different methods of estimating geodesic distance from the original inter-document proximities. Classical Isomap is not parameter free, as geodesic distance is computed by traversing a K-nearest neighbour graph. The experiments first sought to determine the impact of selecting a minimal K-graph, which can be determined automatically, and an optimal K-graph which requires experimentation

and prior knowledge of expected topic clusters. Two alternative pruning algorithms, namely minimum spanning tree and pathfinder network scaling (Schvaneveldt et al., 1989) were also compared. It was found that whilst both algorithms provide viable alternatives to K-neighbours, strong results from minimum spanning trees are particularly interesting, suggesting a simple and reliable solution to the K-Isomap parameter problem.

The rest of the paper is arranged as follows. In the next section we explain the rationale for using the Isomap along and discuss related evidence which supports the use of minimum-spanning tree and pathfinder as alternative graph pruning algorithms. Two key research questions are then defined. In section three, the methodology for the two experiments is described. The results of experiment one and two are reported and discussed in sections four and five respectively. Finally, section six draws conclusions and discusses the implications for future work.

2. Spatialisation using geodesic distances

Spatialisation has been applied extensively by visualisation researchers to many problems that involve exploration of a high-dimensional data space. The distance-similarity metaphor is desirable because it exploits our innate tendency to form a correspondence between perceived spatial proximity of objects and their conceptual similarity (Tobler, 1970; Montello et al., 2003). Spatialisation of document spaces became a popular research field during the 1990s (Chalmers and Chitson, 1992; Wise et al., 1995; Lin, 1997). Empirical studies conducted since then have indicated that users are able to effectively exploit distance-similarity structures during a variety of document (Hornbaek and Frojaer, 1999; Allan et al., 2001; Westerman et al., 2005; Butavicius and Lee, 2007) and conceptual (Westerman and Cribbin, 2000) search tasks. Spatialisations are suitable for a

range of document browsing tasks, ranging from gaining an overview of a large collection (e.g. Wise et al., 1995; Lin, 1997; Chen, 1998; Skupin, 2002) to more directed, local navigation tasks like locating relevant items in a subset of query results (Allan et al., 2001; Cribbin and Chen, 2001). It is the latter task type that this paper focuses on.

The *cluster hypothesis* in information retrieval states that similar documents tend to be relevant to the same requests (van Rijsbergen, 1979). Hearst and Pederson's (1996) seminal paper showed that the cluster hypothesis tends to hold particularly well for documents within a search results set. They found that a clustering algorithm was able to consistently group the majority of relevant items into a single cluster using pair-wise document proximities computed from feature (i.e. word term frequency) vectors. By focusing only on this optimal cluster, the user is able improve precision (and thus speed) of their search. More recent studies have also confirmed the repeatability (Tombros et al., 2002) and usability (Wu et al., 2001) of this dynamic or query-specific clustering approach.

However, a clustering approach like Scatter/Gather (Hearst and Pederson, 1996) only partitions the list of retrieved documents into several smaller, albeit more homogeneous lists. Leuski (2001) observed that this approach is not an optimal solution, for two key reasons. First, no direct cues to inter-document similarity are provided within clusters. As even the best cluster will tend to contain a significant number of non-relevant items, cues to relative proximity are likely to be useful and will increase precision still further. Second, no cues to similarity are provided for documents in different clusters, hence the smaller proportion of relevant items that do not make it into the optimal cluster will be difficult to locate without resorting back to an exhaustive search strategy.

Leuski (Allan et al., 2001; Leuski, 2001) suggested that a spatialisation of document proximities could provide all the benefits of discrete clustering (i.e. a single optimal cluster) whilst addressing these shortcomings. The Lighthouse project achieved a successful demonstration of this concept for a large sample of test queries. By searching unseen documents in order of their proximity to one or more known relevant items, users (both real and simulated) were able to locate other relevant items more quickly and efficiently than was possible using a traditional ranked list layout. This will be referred to as the *cluster growing strategy*.

Whilst the concept of the cluster growing strategy is intuitively appealing, its application in real-time interactive search systems is obstructed by the need to fine-tune the spatialisation to the specific characteristics of each new document set. Leuski (2001) and others (e.g. Swan and Allan, 1998; Rorvig and Fitzpatrick, 1998) have found that achieving an acceptable layout is far more complex than directly inputting the proximity matrix into a conventional layout algorithm like multi-dimensional scaling (MDS: Coxon, 1982) or a spring-embedder (see Chen, 2004). These algorithms seek a globally optimal solution which can often lead to visualizations that are rather amorphous in structure, making it difficult to identify clusters (Rorvig and Fitzpatrick, 1998). Moreover, this lack of structure is typically symptomatic of severe compromises made during the layout process. Nodes that should be close to one another become separated, whilst nodes that should be distal become near neighbours in output space. The first error reduces the *continuity* of the spatialisation, whilst the second reduces its *trustworthiness* (Venna and Kaski, 2006). Misplacements that reduce trustworthiness are particularly problematic for our task, given that the main goal is to improve the precision of document navigation.

Hornbaek and Froekjaer (1999) for instance found that users of an MDS derived spatialisation became confused when proximal documents appeared to have little in common. Worse still, sometimes users would wrongly assume certain documents were closely related simply because they were nearby.

A principal reason why global approaches fail is that there is simply too much information (variance) in the proximity graph to represent within the confines of visual space. The intrinsic dimensionality of a document feature space is normally much higher than two or even three dimensions and so misplacements will occur simply because there are too few degrees of freedom available to map the information to output space accurately. The problem can be imagined if one tries to preserve the proximities between points on the surface of a cube within only one or two dimensions.

An obvious solution is to reduce the information that must be preserved by focusing primarily on preserving only local structural detail. Leuski (2001), following the example of Swan and Allan (1998), achieved this by squaring all similarities that fell below some threshold, thus suppressing their value considerably (given similarities in the range of 0 to 1). A spring-embedder (SE) algorithm was used to create the layout rather than conventional MDS. SE treats documents as rings that are connected by springs. The strength of each spring is determined as a function of similarity between the two nodes it connects. Leuski (2001) found that after an optimal threshold transformation documents coalesced into several tight clusters, with one of those clusters typically containing the majority of relevant items. Unfortunately, this approach is highly sensitive to the choice of threshold, which must be optimised manually for each query/document set. Set the threshold too low and the relevant cluster will scatter; too high and cluster becomes

polluted with non-relevant items. Swan and Allan (1998) got around this by selecting the initial threshold arbitrarily, then allowing the user to adjust it interactively. However, this seems a clumsy approach which is likely to impair usability to a lesser or greater extent.

Another key issue with the above approach is that it sacrifices virtually all continuity for the sake of trustworthiness. Whilst trustworthiness might be the priority, the quest for a layout method that is reasonably parameter-tolerant requires an approach that pays at least some heed to continuity during the inevitable trade-off. In this respect, a more recent non-linear dimension reduction method known as isometric feature mapping (Isomap: Tenenbaum et al., 2000) is particularly interesting.

Isomap is similar to the last approach in that it too places an emphasis on preserving the strongest similarities. It too adopts a threshold criterion to transform computed dissimilarities. However, rather than applying an arbitrary suppression function to the dissimilarities that fall above the threshold, they are pruned from the graph completely. A complete graph/matrix is then restored by re-computing all pair-wise dissimilarities as the length of the shortest path or geodesic distance through the pruned graph. A second difference is that the final layout is achieved using MDS, rather than a SE, which takes the matrix of geodesic distances as its input.

Re-computing non-local relationships as a series of short-hops through a connected neighbourhood graph ensures a degree of global continuity whilst maintaining an emphasis on local trustworthiness. Isomap is classified as a non-linear dimension reduction algorithm in that it assumes that the only reliable proximities that can be computed directly from the feature space are those between closely related objects. This is

important, given what is known about the distribution of computed document proximities. Many document pairs will have few or even no features in common whilst a few closely related documents will tend to have many features in common. This leads to a strong skew in the distribution of similarities where the mode is close to or equal to zero and there is a long tail containing a small number of much higher similarities (Muresan and Harper, 2004). MDS algorithms struggle to find meaningful solutions with distributions like this (Martin-Merino and Munoz, 2004). However, after geodesic transformation MDS is able to produce meaningful visualisations of even the most challenging datasets (Tenenbaum et al., 2000). In fact Venna and Kaski (2006) go so far as to say that geodesic distance may be the only true metric of a high-dimensional feature space.

Despite calls for its evaluation (Skupin and Fabrikant, 2003) there have been surprisingly few attempts to evaluate the efficacy of Isomap for document spatialisation (although see Navarro and Lee, 2001; Butavicius and Lee, 2007). Navarro and Lee (2001) examined the ability of Isomap to partition multiple topics within a document set. They showed that Isomap produced spatialisations that were at least as good as MDS, with particular benefits becoming apparent under noisy data conditions.

Navarro and Lee's (2001) analysis was quite limited in that although they compared the effect of Isomap, and different pruning thresholds, on the global level of preserved variance, they did not quantitatively measure the extent to which each method partitioned the known clusters. This paper seeks to quantify the relatively efficacy of Isomap over MDS for the purpose of clustering the intended topic within a query result set. To this end, this paper assumes Leuski's (Leuski, 2001; Allan et al., 2001) cluster growing strategy and

applies a variation on Voorhees (1985) nearest-neighbour test to evaluate the viability of this strategy within a given spatialisation.

Whilst Isomap theoretically provides a better balance between trustworthiness and continuity, it is important to remember that its performance is still highly dependent upon the specification of an optimal pruning threshold parameter (Tenenbaum et al., 2000; Navarro and Lee, 2001). If a K-nearest neighbours criterion is adopted, then the minimal threshold is the lowest K that results in a connected graph. In this paper a minimal criterion, K_{min}, which can be automatically determined, will be compared to an optimal K-threshold, K_{opt}, which can only be determined experimentally.

Although K_{min} can be established automatically there is no guarantee that this threshold will be optimal or even close to optimal (see Navarro and Lee, 2001). Also finding K_{min} may still require substantial searching of the input graph. A second goal of the experiments reported here is, therefore, to investigate a promising alternative graph pruning method that may resolve the parameter problem for Isomap. Chaomei Chen (Chen, 1998; Chen, 2004) is a well known advocate of the use of a geodesic approach to pruning proximity graphs, called Pathfinder Network Scaling (PF: Schvaneveldt et al., 1989). PF simplifies the graph by eliminating links that violate the metric condition of triangle inequality i.e. if the path between two documents can be more efficiently represented as a walk through intermediate documents then the direct link is removed. PF requires two parameters, r and q. The r-parameter defines how shortest-path distances are calculated, specifying the exponent of the Minkowski distance metric i.e. r=1 means that path length is simply the sum of intermediate link lengths, whereas r=2 is the Euclidean distance. The q-parameter specifies the maximum path length for which the triangle

inequality condition must be maintained, from a minimum of q=2 to a maximum of q=N-1.

Despite these two required parameters, it may not be necessary to view them as variables. There is a general consensus in the literature that the most informative spatialisations are produced when the triangle inequality is maintained across the whole of the graph (q=N-1) and r=∞ (Chen, 2004; Quirin et al., 2007). Setting r to infinity essentially means that the length of the direct link is compared only to the longest intermediate link. This configuration typically results in a minimum-spanning tree (MST), containing only N-1 links or, when more than one unique MST exists, PF will find the union of these trees containing just a few more links. This means that even the sparsest PF graph is guaranteed to be connected, another advantage over the threshold pruning criterion.

These so-called minimum-cost PFs have been shown to produce highly legible and usable document spatialisations (Chen, 1999; Cribbin and Chen, 2001; Chen et al., 2002) although spring-embedder algorithms (e.g. Kamada and Kawai, 1989) rather than MDS (i.e. Isomap) have always been used in the past to create these layouts. Whilst Kamada and Kawai's (1989) algorithm works in a similar way to Isomap, considering optimal distance between unconnected nodes to be the geodesic distance, as far as the author knows, no study has yet explored the use of PF (or MST) as an alternative basis from which to estimate geodesic distance within the Isomap procedure.

Despite the fact that PF (q=N-1) finds the set union of MSTs and that studies such as Chen and Morris (2003) have concluded that the extra links lead to more informative spatial structures, it was considered important to also include MST as a condition in these

experiments for two reasons. First, MST is simpler to implement and usually faster to compute than even the most efficient implementations of PF(q=N-1) (e.g. Guerrero-Bote et al., 2006; Quirin et al., 2007). Second, when precise continuous variables are used (proximity values in these experiments were rendered to 5 decimal places) the ties that lead to alternative trees are unlikely to occur. When preparing the graphs for these experiments, none of the 16 datasets resulted in a PF (q=N-1) that was different to its MST.

In this paper, therefore, the answers to the following two research questions are sought. First, how well does Isomap preserve the desired balance between trustworthiness and continuity, compared to MDS solutions using the original proximity graph as input? Second, can a minimum-cost geodesic pruning approach provide an acceptable solution to the parameter problem i.e. produce solutions that are as good as or better than solutions generated from K_{opt} nearest neighbour graphs? Two experiments were conducted in order to answer these questions. Sections four and five present and discuss the results of these experiments. First, the methodology used in these experiments is described.

3. Method

Sections four and five discuss the results from two consecutive experiments where the clustering performance of MDS and Isomap was compared. In both cases, different Isomap solutions were computed using a variety of techniques to derive geodesic distance measurements. Whilst the procedures used to generate the visualisations and conduct the analyses were broadly the same, the two experiments differed with respect to the underlying document sets, topics and the relevance criterion used. Section 3.1 describes and contrasts the two datasets. Section 3.2 describes the methods used to generate the

visualisations. Section 3.3 then describes how the quality of the resulting spatialisations was measured and analysed. Any methodological differences between the experiments are highlighted where appropriate.

3.1. Datasets

Each dataset consisted of a group of document sets, each of which pertained to a single known topic. Topics and their associated documents and relevance data were all drawn from the Text Retrieval Conference (TREC) collection. Specifically topics were selected from the ad hoc and interactive tracks, spread evenly across three different conferences (TREC 6,7, & 8: Voorhees and Harman, 1997, 1998, 1999) in order to control for any potential biases in topic specification or relevance evaluation. All documents were retrieved from the Financial Times news article archive (1991-1994, 210158 articles). In each case relevance data provided post-hoc by the TREC topic assessors was used to determine cluster members. Topic relevance was determined using a pooling method in which the top-ranking documents retrieved and submitted by all competing IR systems were pooled together and evaluated for relevance by an assessor. In the case of ad hoc topics, the relevance model is binary – documents are either relevant or not. For the interactive topics, a two-level model is applied, whereby documents are judged either relevant or not to the general topic then, in turn, relevant items are judged either relevant or not to one or more 'aspects' of the topic.

3.1.1. Ad-hoc topics

The document pool for an ad hoc topic provides a good approximation of the average retrieval set as it is the set-union of all top-ranking documents retrieved by a variety of IR systems using the topic descriptions (queries). Naturally, not all documents in a topic pool

are ultimately judged relevant by the assessor. Selected topics were matched, as far as possible for difficulty, with pooled precision levels all falling within the narrow range of 10.4% - 14.4%. For each chosen topic pool, a random sample of N=100 documents was repeatedly selected until precision equalled 12% (12 documents). Hence these scenarios were low precision, of the kind that might be retrieved by a user with a relatively ill-defined query definition. The topics were: T319 "New Fuel Sources"; T321 "Women in Parliaments"; T343 "Police Deaths"; T353 "Antarctica Exploration"; T354 "Journalist's Risks"; T372 "Native American Casino"; T390 "Orphan drugs"; T404 "Ireland, peace talks"; T416 "Three Gorges project"; T449 "Antibiotics Ineffectiveness".

3.1.2. Interactive topics

The TREC interactive track (TREC 6, 7 & 8) was defined as aspect-oriented retrieval task. Specifically users of competing IR systems were provided with a set of topics and required, in turn, to identify as many different relevant instances (aspects) of the specified topics as possible. In other words, users were not required to locate more than one relevant example of an identified aspect. However, once the results of multiple search sessions were pooled, a set of identified aspect descriptions was defined by the assessor and all pooled documents were assigned a relevance judgement for each aspect definition. As such many aspects are represented by several (and sometimes many) different documents, meaning that relevant documents can be classified as belonging to both the general topic cluster and, in most cases, one or more aspect clusters. This provides a useful two-level cluster hierarchy for evaluation purposes.

The method used here to retrieve the documents themselves was slightly different to that of the ad hoc topics. The TREC document pools represent the end result of interactive

(iterative) searching, meaning that a relatively large proportion of items tend to be relevant to the topic in some way. For this experiment it was considered most interesting to see how well the spatialisation techniques could cluster aspects of a topic retrieved using a single, high-recall query i.e. the aim was to create a document space that might be created, using an initial tentative query, where the user aim is to explore a complex topic. Documents were therefore retrieved from the entire Financial Times archive using a simple (one or two) keyword match search. To maximise aspect recall no limit was placed on the size of retrieved sub-set i.e. all retrieved documents were included in the analyses.

Six topics were selected from the interactive track, taking two from each of three conferences (TREC 6,7, & 8). The only selection requirement was a strong tendency for aspects to be represented (in the relevance data) by several, rather than just one or two documents. This step was taken to ensure that it was possible to compute nearest-neighbour scores (see section 3.3) for as many relevant cases as possible. The selected topics and their properties are summarised in Table 1. The size of the document sets varies widely from just N=127 for T347i to N=588 for T446i. The second column details the total number of topically relevant documents retrieved. Recall was above 70% in all cases. The fourth column details the number of different aspects represented by documents that were retrieved. It can be seen that some document sets are more diverse (e.g. T352i) than others (e.g. T387i). The fifth column details the mean number of aspect relations for the relevant documents. A relatively high mean number of aspect relations relative to the ratio of relevant documents to aspects retrieved indicates a greater tendency for clusters to overlap (e.g. T408i), whilst a relatively low number indicates more distinct aspectual structure (e.g. T347i).

Table 1: Summary of interactive topic-document sets

	Retrieved set size	# relevant docs retrieved	Recall	# aspects retrieved	Mean # aspect relations
T307i: New Hydroelectric projects	137	48	.889	21	3.2973
T347i: Wildlife Extinction	127	33	.767	22	2.2500
T352i: British Chunnel impacts	218	87	.967	28	14.5412
T387i: Radioactive waste	162	39	.886	9	26.9474
T408i: Tropical storms	122	53	.736	15	35.5111
T446i: Tourists, violence	588	55	.966	15	12.7347

3.2. Spatialisation procedure

The basis for all semantic models was a word term-document vector space model, using the product of term frequency and inverse logarithm of document frequency (TFIDF) to weight the importance of each term within each document (Salton and McGill, 1983). The vocabulary of terms used to define a given vector space included only terms that were present in the 'retrieved' document sub-set i.e. the vocabulary was query-specific. No stemming was done but stop-words and both frequent (df > N*0.9) and unique (df = 1) terms were removed from the vocabulary to improve discrimination. An inter-document similarity matrix was then computed using the cosine measure, which was then converted to a dissimilarity matrix using the function: *dissimilarity = 1-similarity*. This 'original' semantic model will be referred to as the direct cosine dissimilarity (DCD) graph as all

inter-document proximities represent the direct distances between documents within vector space, as opposed to geodesic distances which are computed by traversing the pruned neighbourhood graph. All the remaining semantic model conditions were derived from this matrix by first pruning links according a specified criterion (see below) and then computing inter-document geodesic distances using a shortest path algorithm.

3.2.1. Nearest neighbour graph pruning

K-nearest neighbour graphs were computed by searching each column of the DCD matrix for the top-K lowest dissimilarity values. If a tie occurred, then the first case encountered was retained over later cases. A K_{min} graph represents the lowest possible value of K that results in a connected graph. A K_{opt} graph was determined by computing all graphs from K=2 to 7. The optimal graph was determined as the one that resulted in the best spatialisation as defined by trustworthiness score (defined below in section 3.3). Tables 2 and 3 detail the links retained in all graphs for the ad hoc and interactive datasets respectively.

3.2.2. Geodesic graph pruning

MST graphs were computed using Prim's (1957) algorithm. The PF graphs were derived using the implementation of Schvaneveldt's (Schvaneveldt et al., 1989) algorithm provided by the InfoVis Cyberinfrastructure (http://iv.slis.indiana.edu/). In line with common practice, infinity was used as the r parameter. However, whilst common practice also suggests the use of N-1 for the q parameter to ensure the union of all MSTs (Chen, 2004), it was found that all such graphs were identical to the MSTs derived using Prim's algorithm. Through experimentation, it was found that q=6 was the highest setting that

provided a small but consistent increase in preserved links over MST. Hence this was adopted as the minimal PF condition (PF6). To explore further the effect of manipulating the q parameter, it was decided to also compute graphs using q=3 (PF3). The PF3 graphs tended to retain considerably more links than PF6 as can be seen in Table 2.

Table 2: Summary of retained links in all pruned graphs for the ad hoc dataset

	MST	PF6	PF3	K_{min}	K_{opt}
T319	99	103	129	145 (K=2)	215 (K=3)
T321	99	103	120	149 (K=2)	149 (K=2)
T343	99	102	140	139 (K=2)	139 (K=2)
T353	99	103	132	206 (K=3)	394 (K=6)
T354	99	102	147	203 (K=3)	325 (K=5)
T372	99	103	132	212 (K=3)	212 (K=3)
T390	99	103	136	145 (K=2)	219 (K=3)
T404	99	100	123	284 (K=4)	496 (K=7)
T416	99	100	131	148 (K=2)	497 (K=7)
T449	99	103	132	213 (K=3)	468 (K=7)

Table 3: Summary of retained links in all pruned graphs for the interactive dataset

	MST	PF6	K_{min}	K_{opt}
T307i	136	138	290 (K=3)	290 (K=3)
T347i	126	136	183 (K=2)	183 (K=2)
T352i	217	223	479 (K=3)	640 (K=4)
T387i	161	165	240 (K=2)	671 (K=6)
T408i	121	125	262 (K=3)	262 (K=3)
T446i	587	625	1301 (K=3)	1301 (K=3)

3.2.3. Spatialisation

All neighbourhood and geodesic graphs were represented as N by N matrices. Geodesic distances were then computed using Floyd's shortest paths algorithm (Floyd, 1962). Spatial layout was achieved by inputting all matrices into PROXSCAL (Busing et al., 1987) as provided by SPSS v.15. Preliminary experiments showed significantly better results when non-metric MDS was used to scale the DCD graph. Proximity transformation model had no significant effect on the quality of Isomap spatialisations. Hence to match the MDS and Isomap conditions, non-metric MDS was chosen for all conditions. Other PROXSCAL settings were left on default, which meant a Simplex initial configuration with stress convergence and minimum stress set to 0.0001 and maximum iterations set to 100.

3.3. Experimental Design

For the remainder of this paper, Isomap solutions based on graphs generated using a nearest-neighbour criterion will be collectively referred to as K-Isomap. Individually the conditions are referred to as Iso_{min} (minimum threshold) and Iso_{opt} (optimal threshold). Isomap solutions based on a graph generated using a geodesic criterion (i.e. MST or PF) are collectively referred to as G-Isomap. Individually the conditions are referred to as Iso_{MST} and Iso_{PFq} (where q refers to the shortest-path threshold specified).

The control condition was the original dissimilarity graph, DCD. The experimental treatment conditions were the spatialisation methods applied to this graph: MDS, Iso_{min}, Iso_{opt}, Iso_{MST}, Iso_{PF3} , Iso_{PF6}. Iso_{PF3} was not included in Experiment two, due to poor performance in the first experiment.

3.3.1. Dependent variables

Dependent measures were computed using an adaptation of the nearest-neighbours cluster hypothesis test (Voorhees, 1985). This test considers only the ordinal distribution of document similarities, making it a better measure of spatialisation quality for cluster growing and local navigation strategies than more quantitative alternatives like the cluster separation test (van Rijsbergen and Sparck-Jones, 1973).

A nearest-neighbour test (NNT) score is computed for each relevant document by counting the proportion of the K most similar documents that are also relevant to the same query (topic or aspect). If one considers each relevant document as a potential exemplar or stimulus for cluster-growing, then it makes sense to transform the NNT into a precision or recall figure by either dividing the score by K for precision, or by the total number of known relevant documents for recall. For the experiments reported, NNT scores were computed at K=5 and K=20. 5-NNT scores were translated into precision scores to give a measure of *trustworthiness* or the likelihood of finding other relevant items within the immediate neighbourhood of a relevant document. 20-NNT scores were translated into recall scores to provide a measure of *continuity* or the likely total amount of effort required to locate all relevant documents using a cluster growing strategy.

3.3.2. Analysis

Two main forms of analysis were conducted on the resulting table of NNT scores. First multiple pair-wise comparisons were made using a conservative Bonferroni correction to control for type 1 error. Second, correlation coefficients were computed using Pearson's r. Computing both inferential and correlation statistics provides two different perspectives on the quality of the spatialisations. A relatively high mean trustworthiness score indicates

a tendency for relevant documents to organise locally into more concentrated clusters, whilst a relatively low mean indicates a tendency for localities to be more polluted with unrelated items. Likewise, a relatively high mean continuity score indicates that the known clusters tend to be more cohesive, whilst a relatively low mean indicates a greater tendency to fragment.

Correlations, on the other hand, ignore differences in the means per se, looking instead at the extent to scores tend to co-vary i.e. the extent to which documents that are central (or peripheral) to the cluster in one representation, remain so in another. Hence, they will tell us about the consistency with which individual relevant documents exhibit the same clustering tendencies or 'behaviour' from one representation to the next. For example, a deficit in mean continuity seen in a spatialisation compared to DCD shows some unwanted fragmentation has occurred. However, this might be a result of one or two documents, which were originally fairly central to the main cluster, becoming isolated in the output space. This would suppress the mean considerably. If, however, the general order of 'centrality' of individual documents has remained more or less the same then there will be a strong correlation in scores and we can conclude that the structural fidelity remains good. This approach is particularly important when comparing spatialisations to DCD, although inter-spatialisation correlations can also give us a good indication of differences that occur in the way that compromises are handled during the layout process.

In summary, if correlations with DCD are strong and there is no significant difference in scores or, conversely, correlations are weak and differences highly significant, then clear conclusions can be drawn. However, the results reported next show how discrepancies between the two outcomes can indicate a deficiency in the relevance criterion used to

define the clusters and how selecting a more appropriate criterion can lead to far more insightful results.

4. Experiment one: Topic level clustering

The ad hoc datasets represent a simulation of document retrieval for ten different queries or topics. In each case, documents are either relevant or non-relevant with relevant documents forming 12% of all documents. This gives us a sample size, N, of 120 (12 x 10) cases. The analysis is broken down into parts, addressing the two research questions in turn. Hence K-Isomaps are first compared to MDS and then the relative performance of the G-Isomap solutions is examined. Each part comprises both a comparison of mean trustworthiness and continuity scores, followed by an examination of correlation coefficients to assess degree of differences in structure and structural fidelity (against DCD). First, however, the distributions of computed dissimilarities for DCD, threshold (K_{opt}) and geodesic (MST) pruning are compared.

4.1. Proximity distributions in graphs

Figure 1 shows the effect that that geodesic transformation has on the distribution of proximities in the computed graphs. The expected skewed distribution can be seen in DCD. In fact the modal proximity score is 1.0 (i.e. no similarity). The geodesic graphs show quite different distributions. Although still slightly skewed, in the opposite direction (towards zero), overall these graphs are far more normal in their distributions suggesting an input space that is far more amenable to geometric projection. The pruning criterion has a noticeable effect though, with the MST graphs showing a distinctly smoother

distribution compared to the 'spiky' multi-model distribution of K_{opt}, although this may reflect, in part, the small variation in optimal K from one document set to the next.

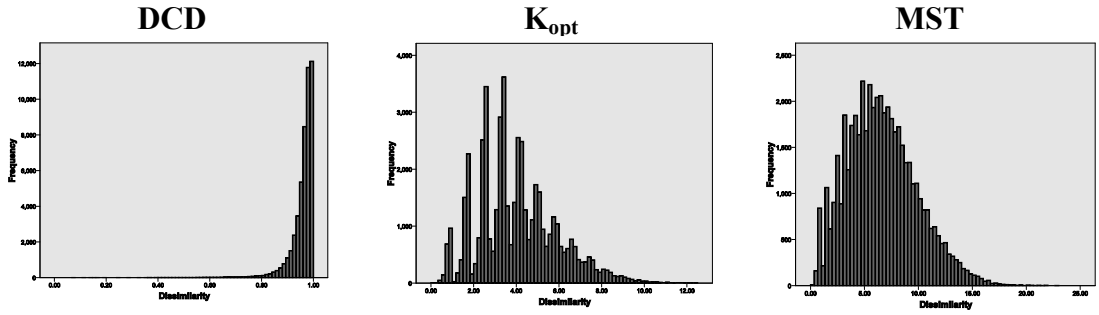

Figure 1: Proximity distributions for graphs. Kopt and MST represent geodesic proximities (all topics combined: n=49500)

4.2. K-Isomap versus MDS

The first stage of the analysis addresses question one by comparing the performance of the standard K-Isomap procedure against MDS. Figure 2 visualises the results, with conditions ordered by mean trustworthiness.

Iso_{opt} provides the most trustworthy solutions overall and, along with MDS, is the only spatialisation method not to differ significantly different from DCD. However, whilst there is a noticeable difference in means between Iso_{opt} and MDS, this difference falls just short of significance ($p=0.10$). Iso_{min} shows comparable trustworthiness to MDS, but is significantly worse than Iso_{opt} ($p<.05$). Hence, it seems that K-Isomap can improve trustworthiness, but this is highly dependent on selection of an optimal threshold. In terms of continuity, MDS is clearly the better method, curiously achieving a slightly higher (but not significantly so) mean than DCD. Iso_{opt} shows significantly ($p<0.05$) poorer continuity than MDS, whilst Iso_{min} is significantly poorer than both MDS and Iso_{opt}.

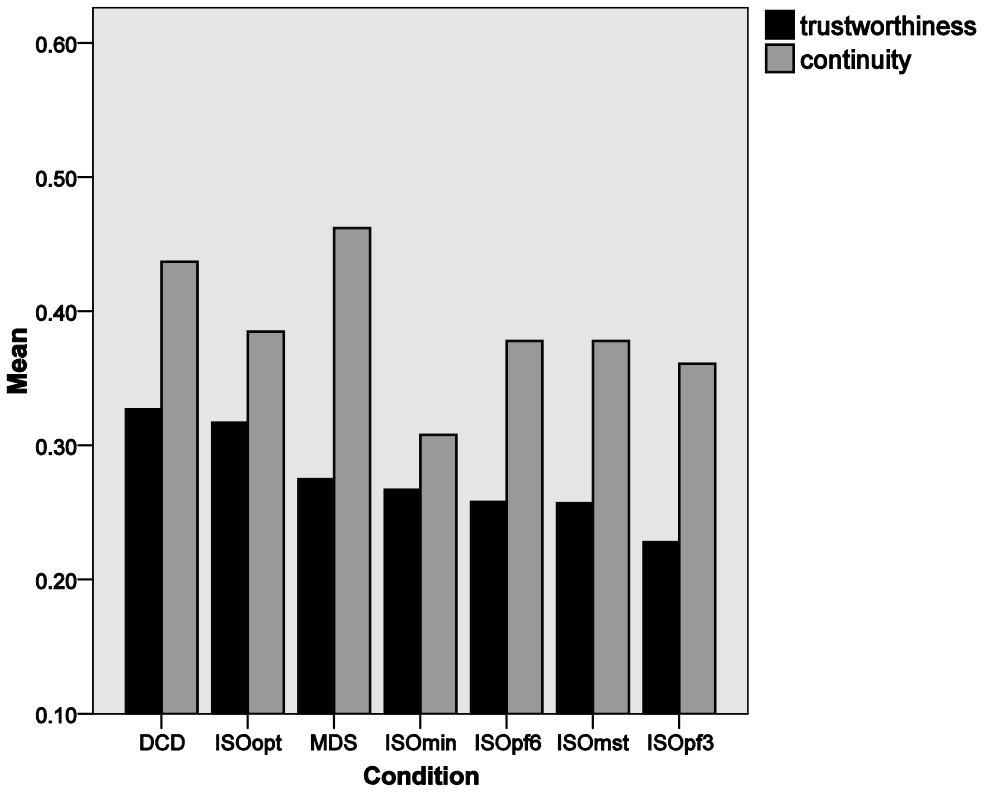

Figure 2: Comparison of mean trustworthiness (topic level precision at K=5) and mean continuity (topic level recall at K=20) by DCD and spatialisation condition for all relevant cases in the ad hoc document sets (n=120)

Table 4 summarises the coefficients computed between spatialisations and DCD, to measure structural fidelity, and between spatialisations to measure structural consistency. Whilst all coefficients are highly significant, they vary widely in magnitude. It is perhaps more meaningful to consider the level of shared variance (r^2) rather than the actual correlation coefficients. There is little difference in the fidelity of the two K-Isomap solutions when it comes to trustworthiness. Both preserve around 50% of the original variance within the DCD graphs, whilst MDS only preserves around 16%. MDS and the K-Isomap solutions share very little variance (~10%) suggesting the approaches represent local structural detail in very different ways. In terms of continuity, K-Isomap and MDS share about the same level of variance with DCD (~40%) but, again, share relatively little

in common with each other (~25%). Iso_{opt} shows some, albeit small advantage over Iso_{min}, reinforcing the importance of selecting an optimal threshold for pruning.

Table 4: Summary of correlations of trustworthiness (precision at K=5) and continuity (recall at K=20) for aspect relevance (d.f. = 118). All coefficients are significant ($p<0.01$) except where indicated (n.s.).

	Trustworthiness (P @ K=5)				Continuity (R @ K=20)			
	DCD	MDS	Iso_{min}	Iso_{opt}	DCD	MDS	Iso_{min}	Iso_{opt}
MDS	0.406				0.641			
Iso_{min}	0.695	0.310			0.626	0.459		
Iso_{opt}	0.722	0.345	0.825		0.680	0.516	0.690	
Iso_{MST}	0.589	0.285	0.578	0.529	0.586	0.494	0.574	0.765
Iso_{PF6}	0.628	0.264	0.672	0.625	0.532	0.440	0.585	0.752
Iso_{PF3}	0.504	0.130 n.s.	0.522	0.490	0.517	0.486	0.463	0.687

To summarise, it seems that K-Isomap can result in some improvements in mean trustworthiness over MDS, but the differences observed in this data are not statistically significant. Moreover, any benefit seems highly dependent upon selecting the optimal threshold. However, K-Isomap is generally better than MDS at preserving local structural fidelity, regardless of pruning threshold. In contrast, K-Isomap performs relatively poorly when it comes to continuity, with significantly lower means in all cases and more or less equivalent levels of structural fidelity. The implications of this discordant relationship between mean comparisons and correlations are discussed later in section 4.4. The next section presents the results of the analysis using the other pruning algorithms.

4.3. G-Isomap

The second stage of the analysis addresses question two, by comparing the relative performance of the G-Isomap solutions. In particular it is important to determine whether the parameter free MST graphs provide an acceptable alternative basis for geodesic distance estimation to graphs pruned using K-nearest neighbour or PF (q<N-1) criteria.

Looking once again at Figure 2, the first thing to notice is that the few extra links retained by PF6 do not result in any material advantage over MST on either measure. Moreover, reducing the q parameter further (PF3) seems only to hurt trustworthiness, although it has somewhat less impact on continuity. In terms of trustworthiness, all of the G-Isomap means were significantly lower than DCD but not significantly different to either MDS or Iso_{min}. More interestingly, although means were a little lower in all cases, only Iso_{PF3} was significantly poorer than Iso_{opt}. Hence, for the sample studied here, an MST or PF6 pruning criterion is certainly as good as a minimum K threshold and, more importantly, not consistently worse than an optimal K criterion.

The advantages of the geodesic pruning criteria become more apparent when continuity is considered, with Iso_{MST} and Iso_{PF6} showing virtually identical means to Iso_{opt} and significantly better means than Iso_{min}. Clearly a focus on preserving shortest paths through the original proximities allows for a better trade-off between trustworthiness and continuity despite the more severe degree of pruning. In fact Iso_{PF6} is the only Isomap method to show no significant difference to DCD, although it could not match the strong performance of MDS, in this respect, being significantly lower.

Turning again to the correlations in Table 4, it is clear that G-Isomap results in some compromises in structural fidelity for both trustworthiness and continuity. At K=5, the fidelity of both Iso_{MST} and Iso_{PF6} is much better than MDS, but somewhat poorer than K-Isomap. At K=20, the G-Isomap solutions are somewhat poorer than all other spatialisations. Once again, it can be seen that adopting a less severe geodesic pruning criterion (Iso_{PF3}) not only reduces mean scores but also hurts fidelity, especially locally (K=5).

It was noted earlier that the K-Isomap solutions shared very little variance with MDS, particularly when it comes to trustworthiness. The local structures of G-Isomap spatialisations seem even more distinct to those of MDS, with Iso_{MST} and Iso_{PF6} sharing only ~7% of variance in trustworthiness and Iso_{PF3} sharing almost no variance (less than 2%). Moreover, the local structures of G-Isomaps are also somewhat distinct from both K-Isomaps. Shared variances are within the range 24% - 45%, with Iso_{PF6} being most similar and Iso_{PF3} being least similar. For continuity, a different picture emerges, with G-Isomap solutions sharing around 50% of the variance with Iso_{opt} spatialisations.

4.4. Discussion

Question one asked whether Isomap can provide more trustworthy spatialisations, whilst maintaining an acceptable level of continuity. Using TREC ad hoc topic relevance as the cluster definition, the analysis presented so far suggests that if an optimised K-nearest neighbour criterion is specified, the Isomap method results in solutions that are slightly, but not quite significantly, more trustworthy than those produced using MDS. All other pruning methods resulted in no significant difference. Moreover, all Isomap methods, including Iso_{opt} resulted in big penalties in terms of continuity relative to MDS.

In isolation these results might suggest that geodesic transformation results in no obvious advantage. However, correlations show that all Isomap methods show much better local structural fidelity than MDS, which sacrifices around 84% of the original variance in DCD scores. Coupled with the poor mean continuity results and similar (K-Isomap) to low (G-Isomap) continuity correlation results, it seems that in the effort to preserve the most local relationships, Isomap has a tendency to disperse the topic cluster. It is possible to see a dispersion effect when the Isomap and MDS spatialisations are compared visually. Figure 3 clearly shows an example of this with the T390 document set. The Iso_{min} and Iso_{MST} solutions clearly present the relevant documents (solid black nodes) as one large main cluster and either one or two smaller clusters, whilst the same documents in the MDS solution form more of a single, loose cluster. The question is whether this dispersion is semantically meaningful.

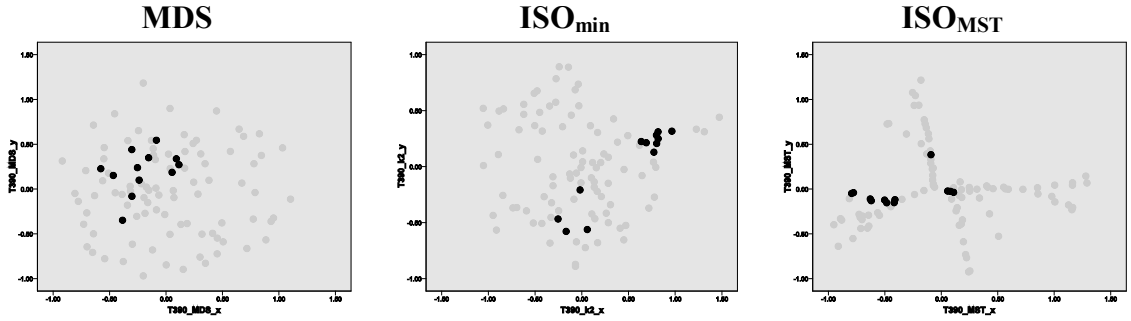

Figure 3: Spatialisations produced for T390. The Isomap solutions clearly show two distinct clusters (black nodes)

Muresan and Harper (2004) observed that it is wrong to assume that documents relevant to the same topic should always form a single, homogeneous cluster. In other words, whilst it may be right to always assume that similar documents are relevant to the same topic, the reverse of this assumption, that all relevant documents are similar, cannot be assumed.

This means there may be more than one relevant cluster each reflecting a different aspect or sub-topic of the overall topic. To account for this Muresan and Harper (2004) revised the cluster hypothesis to the form of the *aspectual cluster hypothesis*:

"Similar documents tend to be relevant to the same requests, but documents relevant to the same requests are not necessarily similar. They tend to be dissimilar if they cover different aspects of the same complex topic" (p.896)

They demonstrated this phenomenon using the TREC interactive topics, where document relevance is categorised according to both topic and, more specifically, known aspects of the topic within the collection. Their analysis showed that most of the strong similarities seen in the tail of the distribution did indeed encode aspect level relationships. Whilst ad hoc topics have been used here, there is no reason to assume that these topics do not comprise multiple aspects. Indeed it is true that many ad hoc topics were recycled to be used within the interactive track. In fact, even casual perusal of the documents used in this experiment quickly revealed strong evidence of aspectual diversity. For instance, in T390 ("Orphan Drugs") there is a clear distinction between eight of the documents, which focus on drugs to treat HIV and AIDS, and the other four which are primarily concerned with the emergence of the bio-technology industry and do not discuss AIDS or HIV at all. The AIDS documents correspond neatly with the larger main cluster seen in the Isomap solutions (Figure 3).

It was hypothesised that the relevance criterion adopted in this experiment was too coarse and may have obscured the value of Isomap because the ad hoc topics might indeed be composed of a number of distinct aspects. This could feasibly account for the disparity in

results between difference tests and correlations. To test this hypothesis, a second experiment was run, this time using TREC interactive topics. It was hypothesised that Isomap would show a much clearer advantage over MDS when aspect level relevance is used as the clustering criterion. The results are presented and discussed in the next section.

Question two asked whether a minimum-cost geodesic pruning criterion could provide the basis for geodesic distance estimation without the need to specify an optimal parameter. The results presented here indicate that an MST or set union of all MSTs (PF, q=N-1) can form the basis of Isomap solutions that are just as trustworthy and continuous as those where geodesic distances are computed from an optimised nearest-neighbour graph. In fact the data suggests that there is little reason to deviate from the minimum-cost criterion, with only very small advantages in trustworthiness and continuity observable for PF when the geodesic criterion is relaxed sufficiently to allow a small number of extra links into the graph (PF6). In contrast, a more severe reduction of the q parameter (q=3) actually hurts the quality of the spatialisation quite measurably. This observation is broadly consistent with the existing consensus that PF is optimal when q=N-1 (Chen, 2004). Hence, in experiment two only MST and PF6 were included as G-Isomap conditions.

5. Experiment two: Aspect level clustering

This experiment followed a similar procedure to the first, but sought to resolve the ambiguities evident in the results of Experiment one by adopting a more specific definition of relevance. Muresan and Harper's (2004) evidence suggests that evaluating the clustering performance using topic relevance as the criterion might have lead to contradictory results because aspectual diversity inherent to the topics was resulting in more than one natural cluster. Visual inspections of the T390 spatialisations demonstrated

that Isomap was better at distinguishing between two quite distinct aspects of the topic. To test this hypothesis more rigorously, this experiment uses a different dataset based on the TREC interactive topics, which allows us to focus on aspect level relevance as the criterion for clustering.

The method used here is broadly the same as for Experiment one. The main difference was in the method used to compose the document sets. All semantic modelling, graph pruning and spatialisation procedures were the same, except for the PF3 pruning condition which was excluded on the basis of its poor performance with the ad hoc datasets.

As in Experiment one, results are presented in two parts. The first focuses on the comparison between K-Isomap and MDS. The second part introduces and compares the G-Isomaps with all other spatialisations. In both parts, a comparison of mean trustworthiness and continuity scores is presented first followed by an examination of correlation coefficients.

5.1. K-Isomap versus MDS

Aspect level cluster analysis reveals quite a different set of trends to the topic level analysis. Figure 4 shows conditions ordered by mean trustworthiness. Inevitably, all spatialisations suffer a significant loss of trustworthiness and continuity in comparison to the original high-dimensional DCD graph. However, K-Isomap consistently performs significantly better than MDS in terms of both trustworthiness and, more significantly, continuity. Whilst there is a slight advantage associated with experimentally selecting the optimal threshold, these differences are not significant. This contrasts sharply with the previous results, where Iso_{opt} performed significantly better than Iso_{min}. One explanation is

that for this experiment, optimal K was higher than minimal K for a much smaller proportion of topics than was the case with the ad hoc sets (33% vs. 60%). Whether this represents a sampling bias or a general tendency for K to be less critical to Isomap for more specific clustering criterion is unclear.

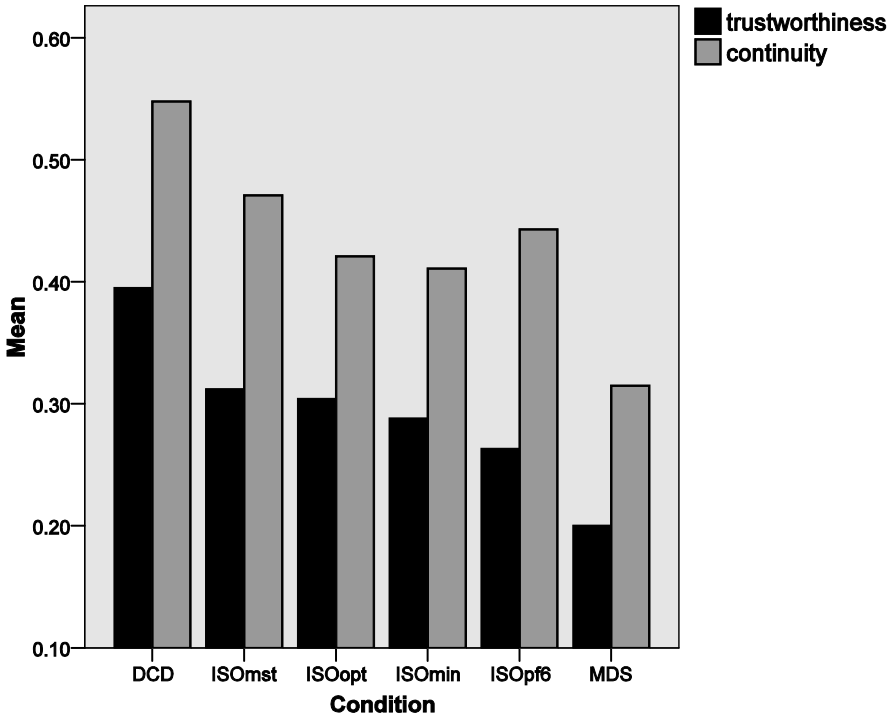

Figure 4: Comparison of mean trustworthiness (aspect level precision at K=5) and mean continuity (aspect level recall at K=20) by DCD and spatialisation condition for all relevant cases in the interactive document sets (n=278)

Table 5 shows the correlations at the aspect relevance level. MDS shares relatively little variance with DCD, even when it comes to continuity. This contrasts sharply with the results in Experiment one. The advantage of Isomap over MDS, however, is reinforced further, with structural fidelity being much higher than MDS for both trustworthiness and continuity scores. Relatively low correlations between the K-Isomaps and MDS emphasise substantial structural differences, particularly at K=20.

Table 5: Summary of correlations of trustworthiness (precision at K=5) and continuity (recall at K=20) for aspect relevance (d.f. = 276). All coefficients are significant ($p<0.01$)

	Trustworthiness (P @ K=5)				Continuity (R @ K=20)			
	DCD	MDS	Iso_{Min}	Iso_{Opt}	DCD	MDS	Iso_{Min}	Iso_{Opt}
MDS	0.402				0.381			
Iso_{min}	0.621	0.440			0.652	0.291		
Iso_{opt}	0.564	0.478	0.774		0.654	0.286	0.955	
Iso_{MST}	0.640	0.486	0.624	0.520	0.794	0.312	0.623	0.609
Iso_{PF6}	0.611	0.489	0.596	0.520	0.813	0.361	0.659	0.640

As seen with the mean comparisons, the benefit of selecting an optimal threshold is less clear than was the case in Experiment one. Correlation coefficients are similar on both measures, with Iso_{min} actually providing slightly better fidelity than Iso_{opt} at K=5. Moreover, the variance shared between Iso_{min} and Iso_{opt} is quite high for trustworthiness (60%) and very high when it comes to continuity (91%). This suggests little overall advantage in fine-tuning the threshold.

5.2. G-Isomap

Despite their much lower link counts the performance of the MST and PF6 algorithms is generally impressive. Both methods result in better solutions than MDS ($p<0.001$) and either equivalent or better spatialisations than both K-Isomaps. As with K-Isomap, the improvement in continuity means are particularly impressive, given the large differences seen in Experiment one. It seems that there is a reduction in the trustworthiness-continuity trade-off for all Isomap methods as the definition of relevance becomes more specific. Iso_{MST} returns the best performance overall, particularly in terms of preserving continuity,

where it is significantly better than Iso_{min} and nearly significantly better (p=.07) than Iso_{opt}. This is an important result, given question two. Iso_{PF6} returns a respectable performance, but shows poor trustworthiness relative to Iso_{MST} ($p<.001$) and, to a lesser extent, the K-Isomaps (not significant). Continuity scores are somewhat poorer than Iso_{MST} (p=0.08) but slightly (not significant) better than the K-Isomaps.

Turning to structural fidelity, Table 5 shows that both G-Isomaps are on a par with K-Isomap in terms of trustworthiness, with all correlations in the region of 0.6 (36% shared variance). Structural fidelity of both Iso_{MST} and Iso_{PF6} at K=20 is considerably higher, with both sharing ~64% of variance with DCD, compared to K-Isomap solutions which only shared ~43%. The variance shared with MDS is similar for both G-Isomaps, being quite low (~24%) for trustworthiness and very low (~10%) for continuity, indicating once again large structural differences between Isomap and MDS. However, the variance shared between G-Isomaps and the K-Isomaps is also relatively low.

5.3. Discussion

The results of Experiment one were somewhat ambiguous. On the one hand, correlations of trustworthiness suggested better local structural fidelity in the Isomap spatialisations, compared to MDS, yet no significant differences in trustworthiness were observed. Moreover, Isomap was shown to be generally poor at preserving continuity. This and visual inspections of the spatialisations suggested that Isomap was causing relevant documents to disperse into more than one cluster. A potential explanation was offered in the form of the aspectual cluster hypothesis (Muresan and Harper, 2004) which predicts that there may be more than one relevant cluster because documents can discuss quite

distinct aspects of the same topic. It was hypothesised that by dispersing the topic, Isomap was in fact preserving and emphasising these aspectual associations and differences.

The results of this second experiment seem to confirm this logical hypothesis. Regardless of the pruning algorithm used, Isomap results in significantly better trustworthiness and continuity than non-metric MDS. In particular, the increase in continuity scores suggests that Isomap is effectively partitioning the different aspects of the topic. This is a significant observation given poor results in a previous attempt to partition aspects using discrete clustering (Wu et al., 2001). Moreover, Leuski (2001) commented on the potential for spatialisation to convey the structure of a complex topic using the results of the "Samual Adams" query, which included documents about both Adams the man and the brand of beer. Although a visualisation showed how such aspect clusters might be conveyed spatially, no algorithmic method of achieving such partitioning was described. Hence, these results have far reaching implications when it comes to developing spatialisation-supported interfaces for exploratory searching. Whilst MDS may have some advantage when it comes to clustering general themes within a document space, Isomap is clearly superior when it comes to preserving the local structures that are needed to support local semantic navigation and exploration.

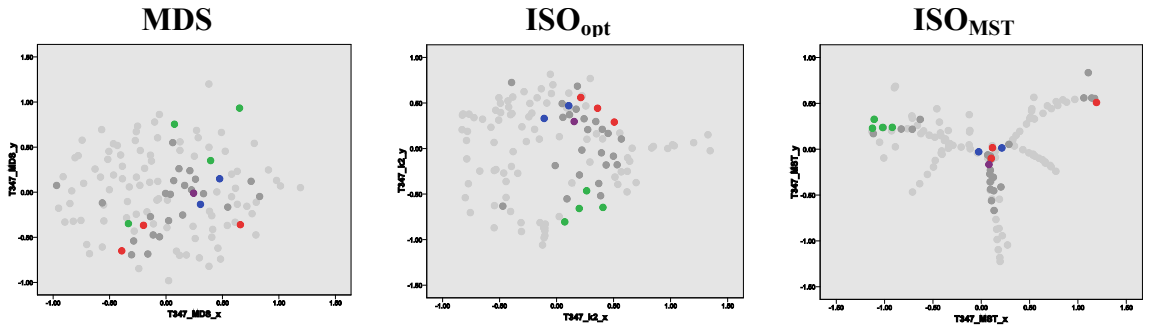

Figure 5: Spatialisations produced for T347. All relevant document nodes are emphasised in red (aspect 7), green (aspect 10), blue (aspect 19), purple (aspects 7 & 19), or dark gray (other aspects). Light gray nodes indicate non-relevant documents

The advantage can be seen by visually inspecting a sample of the spatial solutions. In the T347 document set, the topic separates into relatively distinct (non-overlapping) aspects (see Table 1). Figure 5 shows how these aspects are rendered using MDS, Iso_{opt} and Iso_{MST}. Aspect 10 (green nodes) is completely distinct. Aspects 7 (red nodes) and 19 (blue nodes) overlap slightly, with one shared document (highlighted in purple). As with the ad hoc topic, T390 (Figure 3), whilst MDS seems to separate the general topic (all coloured and dark gray nodes) quite well from non-relevant items (light gray nodes), aspects are not well clustered. Isomap, on the other hand, renders the aspects with greater coherence. What is also evident is how the stricter pruning criterion of MST emphasises aspects 10 and 19 whilst but causes one of the aspect 7 documents to break away into a separate region along with several other topical documents pertaining to other aspects. Iso_{opt}, on the other hand, manages to keep aspect 7 intact but clusters the other aspects less coherently.

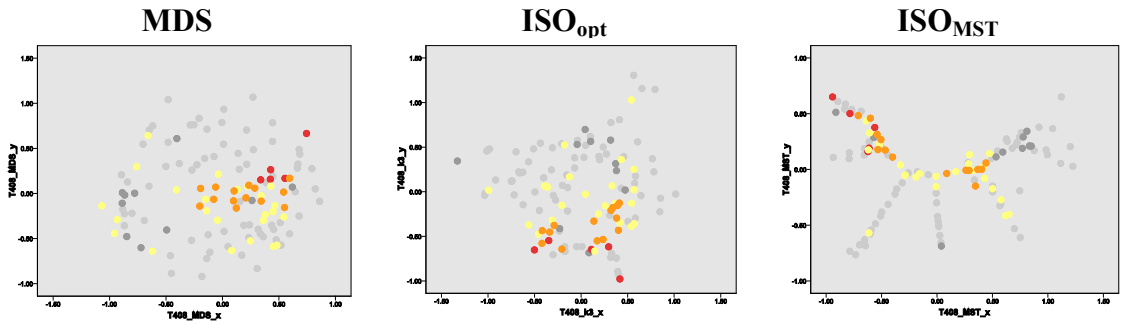

Figure 6 Spatialisations produced for T408. All relevant document nodes are emphasised in red (aspect 2), yellow (aspect 7), orange (aspects 2 & 7), or dark gray (other aspects). Light gray nodes indicate non-relevant documents

T408 (Figure 6) is a more complex, overlapping topic (i.e. documents often relate to more than one aspect). Two aspects (2 and 7) are highlighted in red and yellow respectively. Documents that are relevant to both aspects are highlighted in orange. In this case, the qualitative distinction between MDS and Isomap is less clear. Documents specific to aspect 2 (red nodes) arguably cluster better in MDS than Iso_{opt}, with Iso_{MST} falling somewhere in between. Aspect 7 seems problematic for all algorithms, being distributed quite widely within MDS and Iso_{opt}. The spatial distribution is just as wide in Iso_{MST}, but the tree-like structure provides natural pathways that link many of the aspect 7 documents together, suggesting benefits for semantic navigation. The overlapping documents (orange nodes) are more interesting. Isomap separates these nodes into two clusters. Again this separation is more exaggerated in Iso_{MST} but the 'pathway' structure should pay dividends in terms of supporting user navigation and mental model building (Cribbin and Chen, 2001).

Taking the qualitative and quantitative analyses together, it seems that pruning the proximity graph using the MST algorithm is just as effective if not more so than adopting a K-nearest neighbour criterion, even if K is optimised experimentally. Whilst differences in trustworthiness are small, the biggest benefits appear to be in improved continuity (overall cohesion) of the aspect clusters. The parameter free nature of MST and its algorithmic simplicity make it ideal for real-time applications. Moreover, the minimal edge counts of MSTs translate into very legible and navigable spatialisations (Cribbin and Chen, 2001). Although not experienced during these experiments, there is always the risk of proximity graphs comprising more than one unique MST. However, this possibility can be addressed by substituting MST with more recent optimisations of Pathfinder (Quirin et

al., 2007) which are able to find the set union of MSTs (i.e. PF where q=N-1) with comparable time/space complexity.

The Pathfinder algorithm potentially offered a further advantage over MST in that it is possible to vary, using the q parameter, the link distance up to which the algorithm searches for shortest paths. The results of these experiments show that there is no consistent benefit associated with manipulating q, even to a level that retains just a few extra links. This should be seen as a positive result, given that to achieve any associated benefits would have meant simply replacing the original K-NN parameter problem with a new one.

6. Final conclusions and future work

In this paper, the problem of computing spatialisations, that are sufficiently trustworthy and continuous to support a local foraging or 'cluster growing' strategy (Allan et al., 2001) within a search results set, has been examined. Specifically, the Isomap technique has been compared to MDS, using five different graph pruning methods to generate the required geodesic distances.

The results show that Isomap is particularly effective when it comes to conveying the aspectual structure within a set of search results. This is likely to be particularly useful for exploratory search tasks where the user wishes to browse the results of a high-recall query. Following Leuski's (2001) interaction model, the user might first browse the top-ranked results to identify examples or 'instances' of different aspects of relevance. These instances would then be highlighted within the spatialisation, allowing the user to explore each aspect further by 'growing clusters' from their locations.

Isomap clearly has an important application in the development of spatialisation supported information retrieval systems. However, the role that other non-linear dimension reduction algorithms (e.g. Roweis and Saul, 2000; Venna and Kaski, 2006) might play remains open ground for future evaluation studies. Nevertheless, recent evidence suggests that Isomap provides a relatively desirable balance between trustworthiness and continuity, compared to other non-linear techniques (Venna and Kaski, 2007).

There is strong evidence in the results presented here that the parameter problem that has always been intrinsic to Isomap (Tenenbaum et al., 2000) can be effectively addressed, for document visualisation at least, by a minimum-cost geodesic approach to graph pruning. Our data show that this criterion is particularly effective when the task requires the discrimination of relatively specific features (i.e. aspects as opposed to more general topics). However, the results from Experiment one show that the MST algorithm can provide acceptable solutions even for more coarsely defined topics, without the need for either an automatic or manual search for a connected and/or optimal graph. Future work evaluating the using of minimum-cost graph pruning in other domains, like bioinformatics, seems very worthwhile in this respect.

Clearly it is important to verify the validity of these approaches further by means of formal user studies. These should examine differences not only between MDS and Isomap per se, but also between different graph pruning techniques, as studied here. A recent study by Butavicius and Lee (2007) showed no advantage to users when Isomap was used over MDS. However, it is important to note that only one small document set was studied and proximities were derived using human judgements, rather than from feature vectors.

Hence, the 'curse of dimensionality' issues that cause the dramatic skew which tends to occur in the distribution of computed dissimilarities may not have applied in this case. Elsewhere, Cribbin and Chen (2001) have demonstrated that users searched more effectively when spatialisations were created from graphs pruned using MST and PF. However, this study used a spring-embedder algorithm, rather than MDS, to layout the pruned graphs and did not include K-NN graphs as comparisons.

A final point to make is that the positive results shown by the Isomap technique have highlighted a further avenue of research which might lead to significant reductions in the computational cost of spatialisation. Currently, to compute any document spatialisation it is necessary to first compute proximity values for all document pairs in high-dimensional space. The complexity of this procedure is of the order O(N^2M) where N equals the number of documents and M equals the number of features (or terms) defining the semantic model. As the Isomap approach only requires a connected graph of neighbourhood proximities it would be desirable to find an algorithm that avoids an exhaustive search in M-dimensional space. To this end, a heuristic solution to the problem of constructing a nearest-neighbours graph of inter-document proximities has already been developed and testing is underway to evaluate the efficacy of this approach.

7. References

Allan, J., Leuski, A., Swan, R., & Byrd, D. (2001). Evaluating combinations of ranked lists and visualisations of inter-document similarity. *Information Processing & Management, 37*(3), 435-458.

Busing, F., Commandeur, J., & Heiser, W. (1997). PROXSCAL: a multidimensional

scaling program for individual differences scaling with constraints. In W. Bandilla & F. Faulbaum (Eds.), *Advances in Statistical Software* (Vol. 6, pp. 67-73). Stuttgart: Lucius & Lucius.

Butavicius, M. A., & Lee, M. D. (2007). An empirical evaluation of four data visualization techniques for displaying short news text similarities. *International Journal of Human-Computer Studies, 65*(11), 931-944.

Chalmers, M., & Chitson, P. (1992). *Bead: Explorations in Information Visualisation.* Paper presented at the Fifteenth Annual International ACM SIGIR conference on Research and development in information retrieval, Copenhagen Denmark.

Chen, C. (1998). Bridging the gap: The use of Pathfinder networks in visual navigation. *Journal of Visual Languages and Computing, 9*(3), 267-286.

Chen, C., Cribbin, T., Kuljis, J., & Macredie, R. (2002). Footprints of Information Foragers: Behaviour Semantics of Visual Exploration. *International Journal of Human-Computer Studies, 57*(2), 139-163.

Chen, C., & Morris, S. (2003). *Visualizing Evolving Networks:Minimum Spanning Trees versus Pathfinder Networks.* Paper presented at the IEEE Symposium on Information Visualisation 2003, Seattle, Washington.

Chen, C. (2004). *Information Visualisation: Beyond the Horizon*: Springer.

Coxon, A. (1982). *The User's Guide to Multi-Dimensional Scaling*. London: Heinemann.

Cribbin, T., & Chen, C. (2001). *Visual-Spatial Exploration of Thematic Spaces: A Comparative Study of Three Visualisation Models.* Paper presented at Electronic Imaging 2001: Visual Data Exploration and Analysis VIII, San Jose, CA.

Floyd, R. W. (1962). Algorithm 97: Shortest Path. *Communications of the ACM, 5*(6), 345.

Guerrero-Bote, V. P., Zapico-Alonsoa, F., Espinosa-Calvoa, M. E., Crisóstomoa, R. G., & Moya-Anegón, F. d. (2006). Binary Pathfinder: An improvement to the Pathfinder algorithm. *Information Processing & Management, 42*(6), 1484-1490.

Hearst, M., & Pederson, J. (1996). *Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results.* Paper presented at the 19th Annual International ACM/SIGIR Conference, Zurich, Switzerland.

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).

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters, 31*(1), 7-15.

Leuski, A. (2001). *Interactive Information Organization: Techniques and Evaluation.* Unpublished PhD dissertation, University Of Massachusetts, Amherst.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40-54.

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

Montello, D. R., Fabrikant, S., Ruocco, M., & Middleton, R. S. (2003). Testing the First Law of Cognitive Geography on Point-Display Spatializations, *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 modelling for mediated access to very large document collections. *Journal of the American Society for Information Science and Technology, 55*(10), 892-910.

Navarro, D., & Lee, M. (2001). *Spatial Visualisation of Document Similarity*. Paper presented at the Defence Human Factors Special Interest Group Meeting.

Prim, R. (1957). Shortest Connection Networks and Some Generalizations. *Bell System Technical Journal, 36*, 1389-1401.

Quirin, A., Cordóna, O., Santamaríab, J., Vargas-Quesadac, B., & Moya-Anegón, F. (2007). A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. *Information Processing & Management, 44*(4), 1611-1623.

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

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*, 2323-2326.

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

Schvaneveldt, R., Durso, F., & Dearholt, D. (1989). Network structures in proximity data. In G. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 24, pp. 249-284): Academic Press.

Skupin, A. (2002). A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications, 22*(1), 50-58.

Skupin, A., & Fabrikant, S. (2003). Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization. *Cartography and Geographic Information Science, Transitions in U.S. Cartography and Geographic Information Science, 30*(2), 95-119.

Swan, R., & Allan, J. (1998). *Aspect Windows, 3-D Visualizations and indirect comparisons of Information Retrieval system.* Paper Presented at the 21st Annual

International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia.

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

Tobler, W. (1970). A Computer Model Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234-240.

Tombros, A., Villa, R., & van Rijsbergen, C. J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing & Management, 38*(4), 559-582.

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

van Rijsbergen, C., & Sparck Jones, K. (1973). A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation, 29*, 251-257.

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

Venna, J., & Kaski, S. (2007). Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization, 6*, 139-154.

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.

Westerman, S. J., & Cribbin, T. (2000). Mapping semantic information in virtual space: Dimensions, variance, and individual differences. *International Journal of Human-Computer Studies, 53*(5), 765-788.

Westerman, S. J., Collins, J., & Cribbin, T. (2005). Browsing a document collection represented in two- and three-dimensional virtual information spaces. *International Journal of Human-Computer Studies, 62*(6), 713-736.

Wise Jr, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995). *Visualising the non-visual: Spatial analysis and interaction with information from text documents*. Paper presented at the IEEE Symposium on Information Visualisation (InfoVis '95), New York.

Wu, M., Fuller, M., & Wilkinson, R. (2001). Using clustering and classification approaches in interactive retrieval. *Information Processing & Management, 37*(3), 459-485.