Available online at www.sciencedirect.com

ScienceDirect

Int. J. Human-Computer Studies 65 (2007) 931–944

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

An empirical evaluation of four data visualization techniques for displaying short news text similarities

Marcus A. Butavicius[a,*], Michael D. Lee[b]

[a]Command, Control, Communications and Intelligence Division, Defence Science and Technology Organisation, 203L, PO Box 1500, Edinburgh, SA, 5111, Australia

[b]Department of Cognitive Sciences, University of California, Irvine, CA, USA

Received 6 June 2006; received in revised form 13 April 2007; accepted 5 July 2007

Communicated by J.D. Fekete

Available online 18 July 2007

Abstract

An experiment was conducted comparing user performance on four data visualization techniques—an unstructured display condition consisting of a random one-dimensional (1D) list and three proximity-based representations including a 1D list ranked by a greedy nearest-neighbor algorithm and two 2D spatial visualizations using the ISOMAP layout algorithm and multidimensional scaling (MDS). Eighty-one participants completed an information retrieval task where the visualization techniques were used to display a corpus consisting of 50 short news texts. Human pairwise similarity judgments for this corpus were used to create the three proximity-based displays. Results demonstrated an advantage in accuracy, the number of documents accessed, and, to a lesser extent, subjective confidence in these displays over the Random List condition and in the 2D over the 1D displays. Similar, but smaller, advantages were observed in the MDS display over ISOMAP however none of these pairwise comparisons were statistically significant. A sequential analysis of participant actions in terms of the proximity of document representations accessed provided some explanation for variations in performance between the displays as well as indicating strategic differences in interactions particularly between visualizations of different dimensionality.

Keywords: Data visualization; Multidimensional scaling; ISOMAP; Empirical evaluation; Human–computer interaction

1. Introduction

The human visual processing system is well adapted for processing large amounts of information. Data visualization attempts to exploit this innate ability to allow a user to perceive patterns, such as groups and outliers, within large data sets in a more efficient manner than examination of the raw data. As a result, data visualization offers the potential to assist in a wide-range of activities where large amounts of information need to be processed. Potential applications include not only research in information science (e.g., Rorvig, 1999a, b) but also data mining, information retrieval and intelligence problems in business, police and defence applications (e.g., Navarro and Lee, 2001).

More specifically, data visualization may be considered as an external artefact supporting decision making and therefore a tool that can extend the human analytic capability (Purchase, 1998; Shneiderman, 1998; Ware, 2000). However, many assessments of data visualization techniques have neglected the ability of a user to use the technique in an accurate and efficient manner. Such assessments have either judged the display on aspects that do not necessarily relate to empirical performance such as computational efficiency or aesthetics (e.g., Tractinsky et al., 2000) or on highly domain specific assessments (e.g., Graham et al., 2000) that do not generalize well to other applications (for a more detailed discussion see Lee et al., 2003a). Because of its role in decision support, it has been argued that data visualization techniques need to be

*Corresponding author. Tel.: +61 8 8259 6097; fax: +61 8 8259 6328.
E-mail addresses: marcus.butavicius@dsto.defence.gov.au (M.A. Butavicius), mdlee@uci.edu (M.D. Lee).

1071-5819/$ - see front matter Crown Copyright © 2007 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2007.07.001

designed in line with cognitive and perceptual principles (Lee and Vickers, 1998; Wise, 1999). In addition, the obvious way to evaluate these displays is by examining human performance empirically; that is, testing how people use these displays in a controlled, experimental manner (e.g., Swan and Allan, 1998; Sebrechts et al., 1999; Sutcliffe et al., 2000; Westerman and Cribbin, 2000; Allan et al., 2001; Wu et al., 2001; Butavicius et al., 2003; Lee et al., 2003a, b; Westerman et al., 2005). Such experimentation is crucial given the evidence that objective and subjective assessments often differ in interface evaluations (Frøkjær et al., 2000; Wu et al., 2001; Lee et al., 2003a).

This paper presents findings from an experiment examining human performance using four visualization techniques—a Random List display (which serves as a reference point) and three proximity-based displays constructed from human judgments of inter-document similarities. Theoretically, when humans use such displays they interpret the relationship between documents presented in the display and use this to find documents relevant to an information need. In our experiment, the document representations are represented as icons and it is the proximity of these icons that convey the similarity of the documents to the user. Consider the case of a user searching for a document on the 'African Humanitarian Crisis'. They may begin browsing the visualization by selecting a particular document at random. If this document is unrelated to the topic of 'African Humanitarian Crisis' then they may disregard the other documents nearby this document. Alternatively, if the document is on the topic of interest they may access the other documents that are nearby until they have found the particular document they need. In this way, visualizations of large corpora can reduce the user effort required to find relevant documents.

We were interested in the fundamental question of how the document similarities should be represented to users and were not testing an entire operational system. By testing the effectiveness of the visualization technique in isolation, our approach is similar to that of Morse and Lewis in their use of *defeatured* systems (Morse and Lewis, 1997), where only the basic features of a system are tested, and BASSTEP methodology (Morse et al., 2002), whereby interfaces are tested in stages. However, our technique differs in using a controlled experimental psychology approach to test the visualizations. In addition to testing a fundamental aspect of information presentation, our approach is also relevant to exploratory document analysis, rather than query-based search, where an understanding of the overall trends in the document space is important rather than simply finding answers to the user's pre-conceived questions.

The proximity-based displays used in this experiment are difficult to test objectively using complex data such as document sets because display construction involves two distinct stages. The first stage involves the derivation of distance measures between the individual items in the set and the second involves the representation of these distances in a display. According to Lee and Vickers (1998), the first stage involves consideration of more 'cognitive' principles while the second stage involves more 'perceptual' principles. In the current experiment, the distances between documents were taken from a previous experiment in which 83 participants were asked to provide similarity judgments on all the pairwise comparisons in the document set (Lee et al., 2005). In this way, the 'cognitive' component of the visualization, namely the similarity measures between the documents, may be considered consistent with the cognitive processing of the participants in our experiment. This is supported by Westerman and Cribbin's (2000) finding that the more the spatial mapping of items in a display depended on actual human ratings, the more effective the display was in assisting information retrieval. By modelling the 'cognitive' component of the document visualization, our experiment tested the performance differences between layout algorithms; that is, the 'perceptual' component of the visualization.

A previous study by Westerman et al. (2005) used machine judgments of document similarity judgments in an experiment to test the effectiveness of visualizing document sets. We believe that the use of actual human judgments of document similarity is important in our study given that there is empirical evidence that machine judgments of document similarity can be inadequate surrogates for human judgments when tested on the exact same documents used in the current study (Lee et al., 2005). Firstly, the correlations were lower than expected with the largest correlation being 0.6 between the human ratings and Latent Semantic Analysis (Deerwester et al., 1990). Secondly, there were systematic differences in the judgments of humans and machine. This was expressed in relatively lower machine similarity judgments for document pairs that were judged most similar by humans. This appears to contrast with Westerman et al.'s (submitted) finding that automatic document similarity techniques compare favorably with human judgments. It may be that the usefulness of automated techniques is influenced by the types of documents analysed and this causes the differences in findings between the two studies. By using human judgments in our experiment, we can better isolate the effects of presentation on the effectiveness of the displays because the variation in performance observed will be influenced by the layout of the data and not the validity of the data itself. Without this control, there would be an additional variable affecting performance in the experiment—namely, variation between what the user and the machine considers similar documents. Even the best layout algorithm may yield poor results because the similarities it accurately presents to the user differ from the user's own semantic space and so the final visualization may not assist the user in their task.

Three-dimensional (3D) displays (or the projection of 3D solutions onto a 2D screen known as $2\frac{1}{2}$ D displays) were not used in the current experiment. It has been argued that, while 3D solutions can capture semantic information more accurately than 2D displays, the additional computing and navigational costs may offset this advantage (Leuski and Allen, 2000; Westerman et al., 2005) and this hypothesis has received some empirical support (Westerman and Cribbin, 2000). In fact, there is some evidence that in tasks like the one used in the current experiment whereby participants are required to find specific documents and not all the documents on a particular topic, performance is sometimes worse than in the 2D equivalent (Sebrechts et al., 1999). From an experimental point of view, the additional navigational functionality associated with a 3D display would also introduce a variable associated with only one display and complicate the comparison with alternative displays.

2. Experimental design

2.1. Document set

The document set consisted of 50 documents selected from the Australian Broadcasting Corporation's NewsMail service. These consisted of short text emails about news stories of between 51 and 126 words in length. We used subjective criteria to arrange these articles into a topic taxonomy as depicted in Table 1. Samples from two documents from the 'African Humanitarian Crisis' topic are:

> Document 48 "The United Nations World Food Program estimates that up to 14 million people in seven countries-Malawi, Mozambique, Zambia, Angola, Swaziland, Lesotho and Zimbabwe-face death by starvation unless there is a massive international response. In Malawi, as many as 10,000 people may have already died …"
>
> Document 3 "In Malawi, as in other countries in the region, AIDS is making the effects of the famine much worse. The overall HIV infection rate in Malawi is 19 per cent, but in some areas up to 35 percent of people are infected…"

Table 1
Topic taxonomy and graph labels

Graph label	Full category name	Number of documents
1	Democrat leadership turmoil	3
2	Robert Mugabe	2
3	Abu Nidal	2
4	Iraq-Russian economic ties	2
5	Nigerian stoning	2
6	Tampa/refugee crisis	3
7	African humanitarian crisis	2
8	Environmental summit/research	4
9	Al Qaeda/Bin Laden	4
10	Iraq tensions	3
11	Floods	2
12	Finance	2
*	Miscellaneous	19

The documents in the 'Miscellaneous' category may be considered to be outside of the topic taxonomy because, unlike the articles under the topic headings, these articles were not semantically related to each other.

2.2. Questions

The task of the participants in the experiment was to answer questions about the documents in the set where all the information required to provide the correct response was contained in the documents themselves. That is, the task required finding facts that were clearly stated within the document such as names, places, dates and numbers. Therefore, the participants did not require any background information about the documents nor did they provide any interpretation or analysis of these documents. It should be noted that, in an operational system, many alternative techniques to find such information (e.g., keyword search) would likely be used. In this study we used the questions to test the ability of the user to identify related/unrelated documents in the set, e.g., to identify clusters, the relationships between clusters and outliers in line with our goal of assessing the visual representation of document similarities (and not a complete information retrieval package).

There were five sets of five questions where each set defined a different information retrieval task varying in the number of documents that needed to be accessed and the relationship between such documents according to the topic taxonomy. Response was by four option forced multiple choice (i.e., participants selected one of four answers where none of these answers was a 'Don't know' option). Each question required access to either:

1. One document in the taxonomy.
2. One document outside of the taxonomy (i.e., from the 'Miscellaneous' category).
3. Two documents both belonging to same topic in the taxonomy.
4. Two documents both from outside of the taxonomy (i.e., from the 'Miscellaneous' category).
5. Two documents both from different topics in the taxonomy.

An example of a question and associated response options from Set 3 (i.e., where participants were required to access two documents that were from the same topic in the taxonomy) is:

> How many people have died in Malawi due to malnutrition (A) and what is the overall HIV infection rate in this country (B)?
> Option 1: (A) 12,000 (B) 9 percent.
> Option 2: (A) 10,000, (B) 19 percent.
> Option 3: (A) 10,000 (B) 9 percent.
> Option 4: (A) 12,000 (B) 19 percent.

2.3. Visualizations

Four different visualization techniques were tested. The first was a *Random List*—a 1D list of the documents in a random order. This display is a surrogate for list-based document representation interfaces where there is no ordering according to semantic similarity. The remaining three visualization techniques used the human similarity ratings to find an appropriate display for a human operator (i.e., they may be considered 'structured'). The *Ordered List* presented the documents in a 1D list format where any two documents that were judged to be highly similar were more likely to be consecutive pairs in the list. More specifically, the list was generated using a greedy nearest-neighbor algorithm based on the inter-document similarities (see Appendix A). This *Ordered List* contrasts with document lists ranked according to relevance to query used in other studies (Allan et al, 2001; Wu et al., 2001).

The remaining two visualization techniques, multidimensional scaling (*MDS*) and ISOMAP, also used human similarity judgments to place more similar documents nearer to each other in the display but did so using a 2D layout. Both MDS and ISOMAP work by finding a coordinate pair for each of the documents such that the distances between these points approximate the original distances (i.e., the similarities from the human pairwise judgments). The MDS display used the standard multidimensional scaling layout approach using the Euclidean distance metric (Cox and Cox, 1994; Lee, 1999). MDS algorithms have previously been applied to data visualization, exploration and analysis (e.g., Mao and Jain, 1995; Lowe and Tipping, 1996; Smith, 2000). In addition, multidimensional scaling has some justification as a model of human mental representation (Shepard, 1957, 1987, 1994) and is used by various psychological models (e.g., Nosofsky, 1986; Kruschke, 1992).

ISOMAP is a layout algorithm that is related to, but more sophisticated than, MDS (Tenenbaum et al., 2000). Although not without its limitations (Balasubramanian and Schwartz, 2002), variants of the technique have already proved very successful in certain applications (Donoho and Grimes, submitted). Instead of operating directly on the original distances, a neighborhood graph is constructed based only on local proximities.[1] In so doing, ISOMAP is suited to mapping nonlinear structures in the document space that are invisible to classical MDS. As with the Ordered List, both solutions were optimized for the purpose of this experiment (see Appendix B).

Initial examination of the structured display solutions with respect to the subject taxonomy revealed that, subjectively, all three techniques provided reasonable representations of the topicality in the data set. In the Ordered List solution all documents from the same topic were next to each other for 11 of the 12 topics (see Table 2). In addition, topicality groups were often adjacent to

Table 2
List visualisations with respect to topic taxonomy

Position	Random	Ordered
1	*	*
2	Democrat leadership turmoil	Al Qaeda/Bin Laden
3	Democrat leadership turmoil	Al Qaeda/Bin Laden
4	*	Al Qaeda/Bin Laden
5	*	Al Qaeda/Bin Laden
6	Al Qaeda/Bin Laden	Abu Nidal
7	Tampa/refugee crisis	Abu Nidal
8	Tampa/refugee crisis	Iraq tensions
9	Al Qaeda/Bin Laden	Iraq tensions
10	Environmental summit/ research	Iraq tensions
11	Floods	*
12	Abu Nidal	Iraq-Russian economic ties
13	Iraq tensions	Iraq-Russian economic ties
14	*	*
15	Iraq tensions	*
16	Robert Mugabe	*
17	*	*
18	Iraq tensions	Tampa/refugee crisis
19	*	Tampa/refugee crisis
20	Abu Nidal	Tampa/refugee crisis
21	*	Democrat leadership turmoil
22	Environmental summit/ research	Democrat leadership turmoil
23	Environmental summit/ research	Democrat leadership turmoil
24	Tampa/refugee crisis	*
25	Finance	Finance
26	*	Finance
27	Democrat leadership turmoil	Environmental summit/ research
28	*	Environmental summit/ research
29	Iraq-Russian economic ties	Environmental summit/ research
30	*	Environmental summit/ research
31	Finance	African humanitarian crisis
32	*	Robert Mugabe
33	Robert Mugabe	Robert Mugabe
34	*	*
35	Nigerian stoning	*
36	*	Nigerian stoning
37	Al Qaeda/Bin Laden	Nigerian stoning
38	Iraq-Russian economic ties	African humanitarian crisis
39	*	Floods
40	*	Floods
41	*	*
42	Nigerian stoning	*
43	Environmental summit/ research	*
44	Al Qaeda/Bin Laden	*
45	Floods	*
46	African humanitarian crisis	*
47	*	*
48	*	*
49	African humanitarian crisis	*
50	*	*

[1]ISOMAP generates the neighborhood graph by either of two methods—by connecting each point to all points within a fixed radius ε, or to all of its K nearest neighbors. From this graph, geodesic distances are calculated between all pairs of points on the manifold and classical MDS is then applied to these geodesic distances.

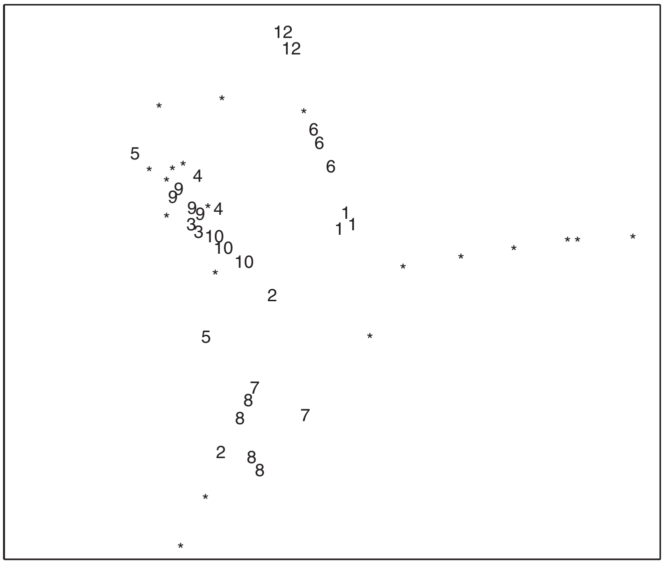

Fig. 1. The two-dimensional solution of ISOMAP labelled according to topic membership (see Table 1 for topic descriptions).

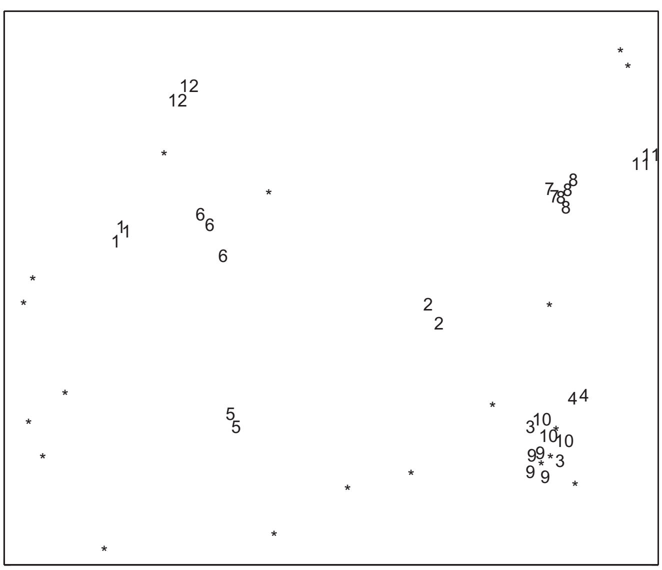

Fig. 2. The two-dimensional solution of MDS labelled according to topic membership (see Table 1 for topic descriptions).

associated topicality groups (e.g., the document cluster on 'Abu Nidal' was adjacent to the cluster on 'Al Qaeda'). It should be noted that the taxonomy was not available to the participants who were providing the pairwise similarity judgments in Lee et al.'s (2005) experiment. In addition, the documents did not have subject headers or titles that might indicate subject matter. However, the topic assignment was subjective and therefore cannot be used in any rigorous test of the effectiveness of the visualization.

Figs. 1 and 2 show the solutions for ISOMAP and MDS, respectively. In these displays, the document representations have been displayed as the graph labels indicating their topicality as specified in Table 2. It should be noted that in the interface these topic labels were not included but

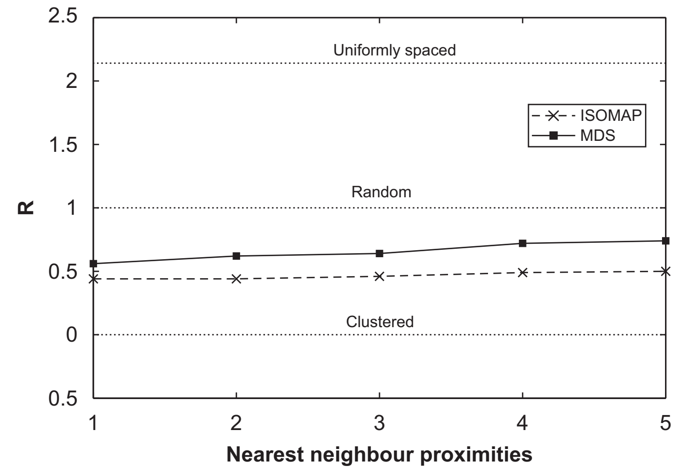

Fig. 3. Values of R for ISOMAP and MDS arrays for 1st through 5th nearest neighbor proximities.

are presented here to demonstrate the way the algorithms displayed the document similarities.

The two 2D solutions, ISOMAP and MDS, appear to provide solutions with different spatial properties. To analyse these differences, we examined the nearest neighbor (NN) statistic, R, for these arrays (for more detail see Preiss, 2006). R provides a quantitative summary of spacing between points and, for a n-NN, is calculated as the ratio of the observed distances (referred to as r_o) to those expected under a random array (r_e) defined by the notion of Complete Spatial Randomness (Diggle, 1983). Values close to 0 indicate clustering, values approaching 1 indicate more randomly distributed points and values approaching 2.149 indicate more uniformly spaced points. For both arrays the values of R indicated clustering at the 1st through to the 5th NN with values ranging from 0.44 to 0.72 (see Fig. 3). Not only were the values of R always higher for MDS, but the degree of clustering reduced with higher NN relations. In contrast, the values of R associated with ISOMAP remained relatively constant across increasing levels of proximity. This indicates that the ISOMAP solution demonstrates greater clustering at all levels of proximities in the arrays. In addition, this clustering was relatively constant across increasing relative proximities of points in contrast to MDS where the amount of clustering dropped off.

2.4. Interface

The interface displaying the ISOMAP solution is shown in Fig. 4. The visualization display is shown in the top-left hand corner. Whenever a point in the display is selected, the related document text is displayed in the top-right hand pane. After a document has been selected, it remains highlighted until the question has been answered to indicate which documents have already been accessed. The question and answer options are displayed at the bottom-left of screen while the confidence scale is shown at the bottom-right. The text of each document fit entirely

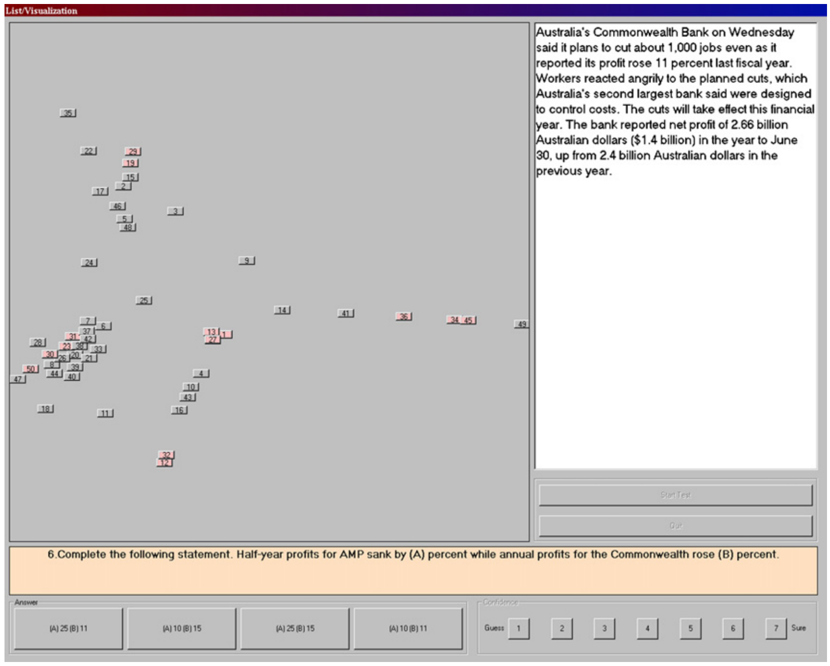

Fig. 4. Screenshot of the experimental interface featuring the ISOMAP display.

within the text pane. In addition, for both of the 1D displays, the entire list of document representations fit on one screen.

2.5. Participants

The majority of the 81 participants in the experiment were students from the Psychology Department at the University of Adelaide. Psychology I students received partial course credit while all other participants received a ten (Australian) dollar gift voucher.

2.6. Procedure

Participants viewed only one of the four visualizations of the document set, however all participants answered the same 25 questions. The order of presentation of these questions was randomized for each participant and the ordering of the four multiple choice options was also randomized for each trial.

After an explanation of the interface was provided by the research assistant, participants were given a practice run using the interface to familiarize themselves with the procedure of the experiment and the operation of the interface. Participants were required to select one of the four answers and then to provide a confidence rating for their judgment on a 7 point scale (where ‘1’ was labelled ‘guess’ and ‘7’ was labelled ‘sure’). Participants were required to provide an answer option before providing the confidence rating and to provide a confidence rating before the next question was presented or, if it was the last question, the experiment was completed.

3. Results

The variable which appeared to best differentiate performance between the four visualization groups was the number of documents accessed. As can be seen in Fig. 5, the distribution of correct values for the unstructured condition is different from those under the structured visualizations. Specifically, there are relatively fewer cases in the Random List condition where less than 40 documents were selected to answer a single question in comparison to the structured display conditions. Fig. 6 shows one standard error about the mean for the same variable for each question across the four conditions. This graph similarly suggests an advantage in the structured visualizations and this result was supported by null hypothesis statistical testing. A repeated measures analysis of variance (RMANOVA) was conducted on the number of documents accessed with a between subjects factor of visualization type (4 levels) and a within subjects factor of question type (5 levels). This demonstrated significant variability in the number of documents accessed between the different visualization types ($F(3,77) = 11.138$, $p < 0.001$, $\eta_p^2 = 0.303$).More specifically, all the comparisons between the Random List and the visualizations based on the human similarity ratings were significant, i.e., fewer documents were accessed using the Ordered List, ISOMAP and MDS displays than the Random List display (Scheffé: $\text{CI}_{95\%}$ $5.281 < \mu_{\text{Random List–Ordered List}} = 33.176 < 61.072$ (SE = 9.759), $p = 0.013$; $\text{CI}_{95\%}$ $15.537 < \mu_{\text{Random List–ISOMAP}} = 43.77 < 72.003$, (SE = 9.878), $p = 0.001$; $\text{CI}_{95\%}$ $25.397 < \mu_{\text{Random List–MDS}} = 53.63 < 81.864$, (SE = 9.878), $p < 0.001$). The biggest advantage was associated with the MDS display where, on average, participants accessed 54 fewer

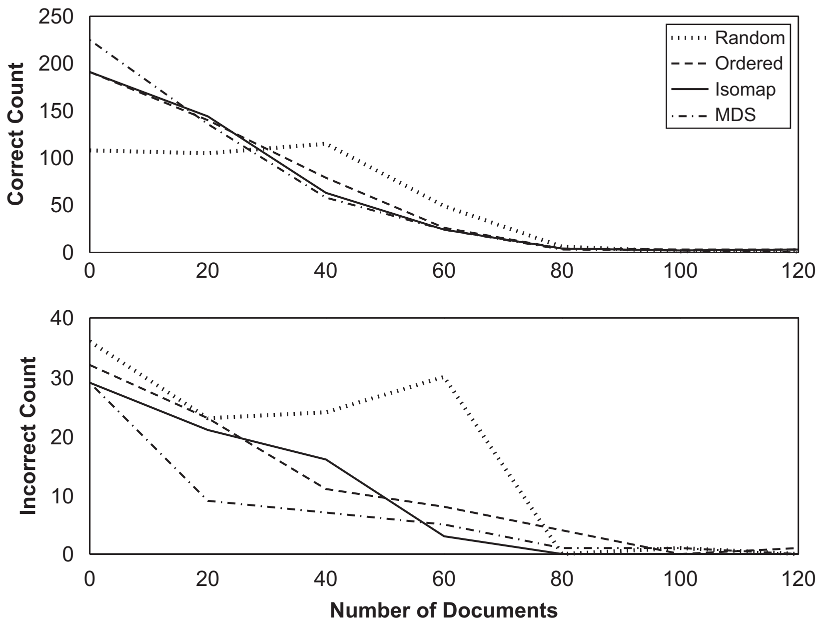

Fig. 5. Distributions of the number of documents accessed for all correct (top) and incorrect (bottom) responses with separate distributions for the four visualization conditions.

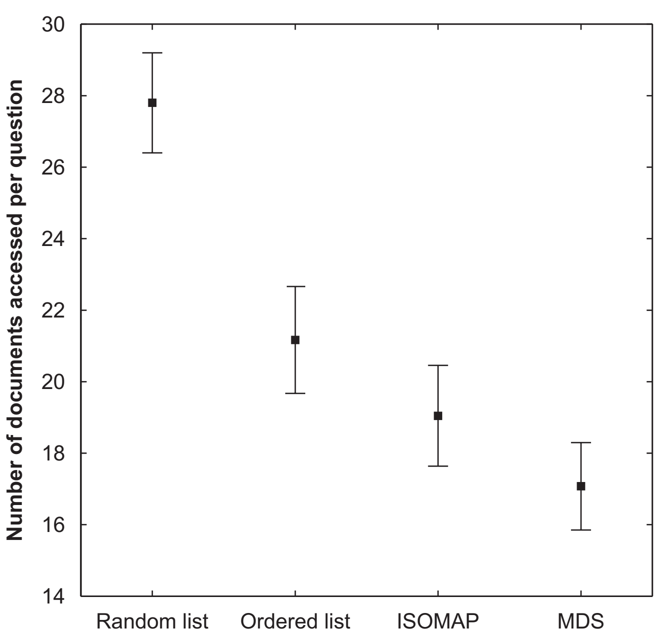

Fig. 6. Number of documents accessed by condition. One standard error is shown about the mean.

documents to answer the questions in comparison to the Random List.

In order to examine more closely the nature of the differences between the visualizations, an analysis of the sequence of documents selected in each trial was also conducted. Of interest were the relative positions of consecutive documents accessed—to what extent were the participants moving from one document to the nearest document rather than making larger jumps in the visualization display? If a visualization is not perceived as being useful then the choice of the next document is arbitrary. Participants may be more likely to minimize the mouse movements by traversing the display via neighboring documents. This is based on the assumption that participants will tend to act in a manner that reduces the amount of effort they expend (Zipf, 1949). In fact, some users may adopt this strategy as a kind of brute force technique to find the relevant document(s) even if they can perceive the structure in the display (but choose not to exploit it). In such cases, however, the outcome is the same because the use of visualization is still not providing any real performance benefit.

To examine the sequence of jumps taken by participants, relative distances between consecutively accessed documents were analysed. A jump between document A and B was considered a NN move when the distance between A and B in the display was the shortest of all the distances involving document A and any other document in the display. Under all four conditions the most frequently occurring jump was between NN documents and overall there was a strong tendency towards smaller jumps-a trend that was greater for the list visualizations (skewness $_{\text{Random List}} = 6.896$, $SE = 0.0214$; skewness $_{\text{Ordered List}} = 5.365$, $SE = 0.0241$) than the 2D displays (skewness $_{\text{ISOMAP}} = 2.636$, $SE = 0.0259$; skewness $_{\text{MDS}} = 2.755$, $SE = 0.0273$). In addition, the percentage of NN moves was considerably higher in the list-based approaches (Random List—85%; Ordered List—80%) than the 2D approaches (ISOMAP–36%; MDS–29%).

Proportions of NN moves were also calculated for each individual and for each question set and a similar

RMANOVA to that performed on the number of documents accessed was conducted. As can be seen in the errorbar graph in Fig. 7, there were differences between the visualization types ($F(3,76) = 167.485$, $p < 0.001$, $\eta_p^2 = 0.869$). In particular, the 2D displays were associated with fewer NN moves than the 1D displays although there was less variability in displays of the same dimensionality (see Table 3). As can be seen in the graph in Fig. 7, there was a relative increase in the proportion of NN moves when the question involved accessing one document outside of the taxonomy (i.e., Question Set 2) for the ISOMAP display in comparison to the other displays and this observation is supported by the significant interaction effect in the RMANOVA (see Table 4). In summary, this analysis appears to demonstrate evidence for strategic differences between users' performance on the different visualizations, particularly between displays of different dimensionality.

There was a similar, albeit weaker, trend for accuracy to that observed in the number of documents accessed. As can be seen in Fig. 8, accuracies under the Random List condition were considerably lower than those for the structured displays. Of the structured displays, the scores for the Ordered List and ISOMAP visualizations were highly similar while those from the MDS approach were highest of all displays. Using a similar RMANOVA to that used to analyse the number of document accessed, the response accuracy varied significantly between the different visualizations ($F(3,77) = 3.061$, $p = 0.033$, $\eta_p^2 = 0.107$). Of the individual comparisons, only the difference between the MDS and Random List displays was significant (Scheffé: $CI_{95\%}$ $-24.51 < \mu_{\text{Random List–MDS}} = -12.4 < -0.291$ (SE = 4.237), $p = 0.043$). On average, participants using the MDS display were 12% more accurate than those using the Random List.

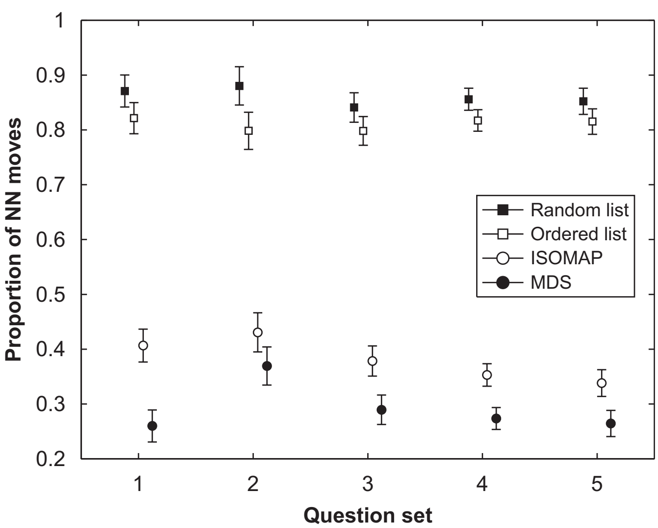

Fig. 7. Proportion of nearest neighbor moves for each question set and condition combination. The question sets are (1) one document in taxonomy, (2) one document outside taxonomy, (3) two documents in same topic, (4) two documents outside taxonomy and (5) two documents from different topics. One standard error is shown about the mean.

For response confidence, the pattern of results was similar to, but weaker than, the trend for accuracy. Examination of the raw data reveals a strong bias towards high confidence responses with 74% of all responses given the highest possible confidence value (i.e., 7 on a scale of 1–7). This trend is evident in Fig. 9. Overall, there was a significant difference in the level of confidence expressed between the four different visualizations ($F(3,77) = 3.286$, $p = 0.025$, $\eta_p^2 = 0.113$). However, while the rank order of the mean scores for the visualizations is similar to that for accuracy, none of the pairwise comparisons between visualizations was statistically significant although the difference between the Random List and MDS displays was close to statistical significance (Scheffé: $CI_{95\%}$$-7.328 < \mu_{\text{Random List–MDS}} = -3.66 < 0.008$ (SE = 1.283), $p = 0.051$).

In contrast, there were no such significant differences in terms of response time. Examination of the raw data reveals less compelling evidence of differences between the conditions (see Fig. 10). As can be seen in Fig. 11, the direction of the differences between means is in fact consistent with the trends for the number of documents accessed, however the overlap between the response time distributions is considerably greater. We examined the hypothesis that the lack of differences in overall time may have been due to participants changing the rate at which they were completing questions in order to finish within the estimated completion time of 45–60 min indicated to the participants before starting the experiment. For example, participants who were slower at the start of the experiment

Table 3
Bonferonni comparisons for the proportion of nearest neighbor moves

Display #1	Display #2	Mean difference$_{\#1-\#2}$ (SE)	P	$CI_{95\%}$
Random list	Ordered list	0.05 (0.031)	0.475	[−0.04,0.14]
	ISOMAP	0.569 (0.032)	<0.001**	[0.477,0.66]
	MDS	0.478 (0.032)	<0.001**	[0.387,0.571]
Ordered list	ISOMAP	0.519 (0.031)	<0.001**	[0.429,0.609]
	MDS	0.429 (0.032)	<0.001**	[0.338, 52]
MDS	ISOMAP	0.09 (0.032)	0.058	[−0.002,0.182]

Note: ** $p < 0.001$.

Table 4
RMANOVA within-subjects effects

Source	df	F	η_p^2	p
		Number of documents		
Question type	4	67.875**	0.469	< 0.001
Question type × Condition	12	.826	0.031	0.607
Question type error	308	(1792.848)		
		Proportion of NN moves		
Question type	4	5.546**	0.068	0.001
Question type × Condition	12	2.342*	0.085	0.014
Question type error	308	(0.006)		
		Accuracy		
Question type	4	15.484**	0.167	< 0.001
Question type × Condition	12	.98	0.037	0.459
Question type error	308	(181.531)		
		Time		
Question type	4	51.476**	0.401	< 0.001
Question type × Condition	12	.973	0.037	0.458
Question type error	308	(4.538×10^{12})		
		Confidence		
Question type	4	28.496**	0.27	< 0.001
Question type × Condition	12	1.351	0.05	0.2
Question type error	308	(10.399)		

Note: Values enclosed in parentheses represent mean square errors. DF values used in calculations adjusted using the Greenhouse–Geisser approach due to violations of the sphericity assumptions in all variables (see Appendix C). *$p<0.05$. ** $p<0.01$.

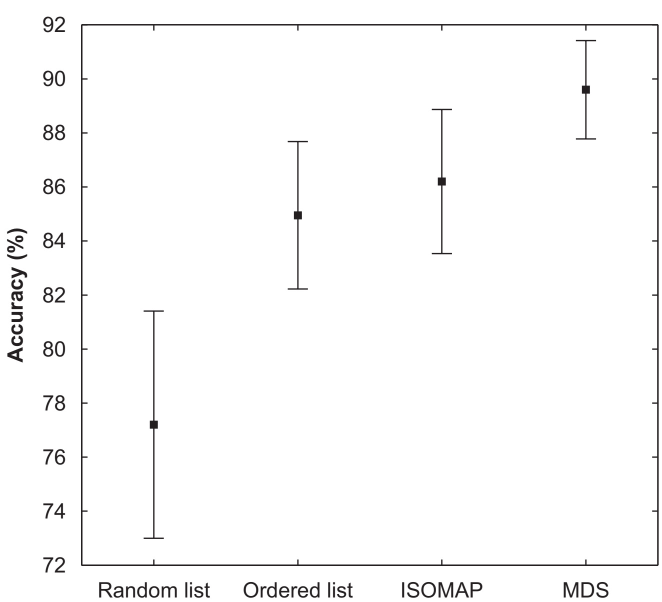

Fig. 8. Accuracy by condition. One standard error is shown about the mean.

may have sped up to finish within the estimated time. However, examination of the cumulative distributions of response times across the four conditions did not reveal any meaningful differences between the visualizations and across the course of the experiment.

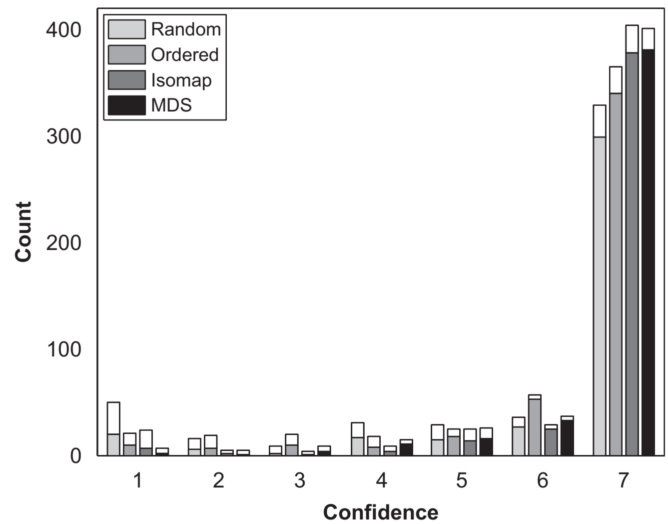

Fig. 9. Histogram of response confidences broken down by visualization and stacked according to accuracy (white area = incorrect responses, colored area = correct responses).

The relationships between the dependent variables were examined using Spearman's coefficient of rank correlation, ρ. Spearman's coefficient of rank correlation, ρ, was used instead of Pearson's product-moment correlation coefficient, r, due to non-normality of the response measure distributions. All correlations were two-tailed. The strongest relationship was the positive correlation between accuracy and confidence ($CI_{95\%}$ $0.666<\rho=0.772<0.847$, $p<0.001$, $N=81$). A similar sized positive correlation was also found between the number of documents accessed and response time ($CI_{95\%}$ $0.346<\rho=0.525<0.667$, $p<0.001$, $N=81$). Negative correlations of smaller magnitudes were also found for the relationships between the number of documents accessed and both confidence ($CI_{95\%}-0.517<\rho=-0.337<-0.128$, $p=0.002$, $N=81$) and accuracy ($CI_{95\%}$ $-0.451<\rho=-0.258<-0.042$, $p=0.02$, $N=81$). Perhaps most interesting were the correlations involving the proportion of NN moves. Larger numbers of NN moves were associated with increasing numbers of documents accessed ($CI_{95\%}$ $0.166<\rho=0.371<0.545$, $p=0.001$, $N=81$) and decreased confidence ($CI_{95\%}$ $-0.482<\rho=-0.295<-0.082$, $p=0.007$, $N=81$) and accuracy ($CI_{95\%}$ $-0.428<\rho=-0.231<-0.013$, $p=0.038$, $N=81$). Analyses of the relationships between the response variables for each visualization were also conducted. The pattern of statistically significant results was generally similar across display types with the exception that there was a unique significant negative correlation between time and confidence for the MDS display ($CI_{95\%}$ $-0.841<\rho=-0.533<0.023$, $p=0.015$, $N=20$) and a lack of any significant correlation between the number of documents accessed and time for the ISOMAP display.

For all the dependent measures (number of documents accessed, proportion of NN moves, accuracy, time and

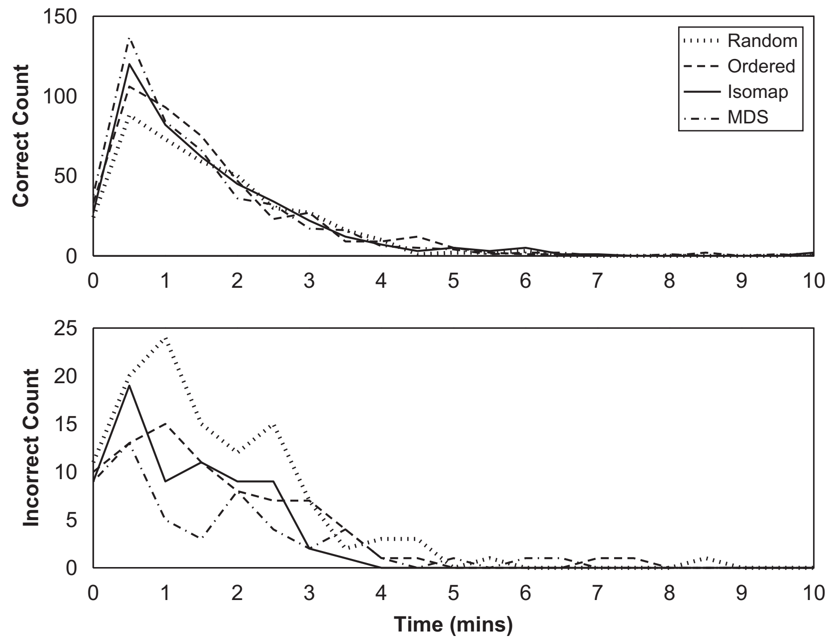

Fig. 10. Distributions of the response times for all correct (top) and incorrect (bottom) responses with separate distributions for the four visualization conditions.

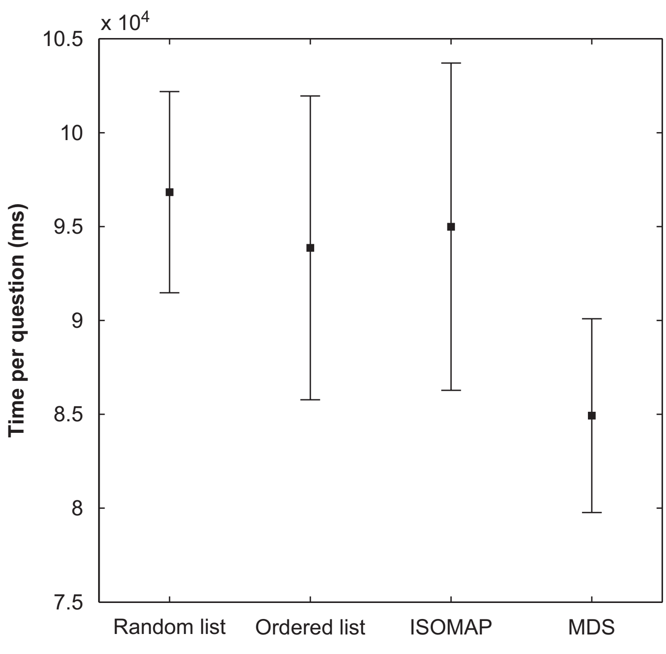

Fig. 11. Time by condition. One standard error is shown about the mean.

confidence) there were significant differences between question type (see Table 4). Overall, the pattern of advantages/disadvantages was similar between the different variables and the best performance was associated with questions that required finding a single outlier document (i.e., a document outside of the taxonomy). However, with the exception of the proportion of NN moves variable, where there was a larger increase on the task of finding an outlier under the ISOMAP condition, the influence of question type did not vary across conditions (i.e., no Visualization by Question Type interaction effects significant at the 0.05 alpha level).

4. Conclusions

In general, this study demonstrated an advantage in the three structured visualization techniques (i.e., those that incorporated human document similarity judgments) over an unstructured display method for presenting documents. This is consistent with previous empirical findings stressing the importance of psychological considerations in the design of visualizations (e.g., Westerman and Cribbin, 2000; Lee et al., 2003a). The best performing structured visualization was MDS followed closely by ISOMAP and these were associated with statistically significant advantages on several dependent variables. The largest effects were observed in terms of the number of documents accessed, with differences in accuracy being next strongest followed by confidence. Improvements in these areas would have important practical benefits in work environments where the accuracy of findings is paramount and when the number of documents available is large. More specifically, there was a significant reduction in the number of documents accessed to answer a question in a list format when it was structured according to human similarity judgments. Of the structured visualizations, the 2D displays were more beneficial than the 1D list. Overall, a slight but consistent benefit in the multidimensional scaling (MDS) approach over ISOMAP was observed over several dependent variables however these differences were not statistically significant.

The difference in performance between the visualizations was not attributable to a particular type of question (for a similar finding see Chen and Dumais, 2000). In other words, the better visualizations helped the user consistently across a range of different tasks from finding two associated documents to finding exceptional documents. The results from this study also support Frøkjær et al.'s (2000) supposition that measures such as speed, accuracy and confidence should be considered independent aspects of usability. In our study the direction of the differences between visualizations, the effect sizes in the RMANOVAs and the correlations varied considerably between the different pairs of dependent variables.

The analysis suggested that strategic differences in the sequence of documents selected may account for some of the differences in performance between the visualizations. The number of NN document representations selected was lower for the 2D versus the 1D displays. In addition, relative increases in the number of times the nearest document was selected were associated with increases in the number of documents selected and decreased accuracy and confidence. Obviously, there may be stages during the search when accessing NN documents is not a brute force strategy but one that relies on interpreting the semantic relationships in the display. For example, if a user has found a document that is on the same topic as the document being searched for, it would be sensible to access documents nearby if the user believed that the visualization had successfully captured the semantic relatedness of the documents in the set. However, a relative increase in the overall number of NN moves in the display may indicate that the user perceives the display to be unhelpful. In such cases, no meaningful search strategy can be found and participants may access neighboring documents as a way of minimizing the effort expended (Zipf, 1949).

The measure of NN moves requires further investigation. In particular, the use of absolute rather than perceived distance in calculating NNs requires investigation. There are at least two reasons why Euclidean distance may not be an appropriate measure to determine relative perceived inter-point distances in the 2D displays. Firstly, in the 2D displays participants may have not perceived marginal differences in distances between points—in psychophysical terms the difference may be below the *differential threshold* (Fechner, 1966). As a result they may have selected the next point from a set of points that, perceptually speaking, are the same distance from the current point but which is not in fact the nearest neighboring point in absolute terms. Secondly, the distances perceived are not necessarily the same as the Euclidean distances in the display (e.g., Levin and Haber, 1993). While there are limitations to this measure, the analysis of user performance on data visualization displays should take into account not just overall dependent variables but sequential analyses of the participants' search strategies. At the very least, it may help to explain *how* some visualizations are better employed by users than others.

An unexpected result of this study was that ISOMAP was slightly inferior to MDS. At first glance, this finding is counter-intuitive given that ISOMAP is better suited to handling more complex document set structures than MDS. However, for the most part the differences between these 2D displays were marginal in comparison to differences either between displays of different dimensionality or structured displays compared to the Random List. It is possible that any observed differences between ISOMAP and MDS are attributable to experimental error. Barring this, the inferior performance of ISOMAP to MDS could be due to one or more reasons. While theoretically, ISOMAP can capture nonlinear structures that are invisible to MDS (Tenenbaum et al., 2000), ISOMAP is not without its limitations and parameter settings must be carefully chosen to prevent the creation of poor solutions where the neighborhood graph generated in the first stage of the algorithm misrepresents the underlying data structure (Balasubramanian and Schwartz, 2002).[2] In addition, comparing the theoretical performance of the two algorithms is not straightforward (see Appendix B). The interaction between display algorithm and document set is also another issue that complicates the assessment of specific visualization techniques. It is possible that the performance of any algorithm is related to data set characteristics such as size, centrality and reciprocity (see Tversky and Hutchinson, 1986). Such issues are worthy of further investigation.

While the underlying distances on which MDS and ISOMAP operated in this experiment may be considered ideal from a cognitive point of view, such algorithms may still neglect certain visual aspects of data visualization tasks. As mentioned previously, perceived distance in the display is likely to differ from Euclidean distance. Therefore, even if either algorithm had accurately captured the representational space of the cognitive similarities between the documents, the user may not faithfully interpret this visual space. In addition to the problem of representing simple inter-point distances, the observed relationships between documents will also be influenced by the perception of local and global structures in the display. The advantage of data visualization is that it exploits the users' perceptual apparatus to quickly determine structures and patterns in the display. However, the perceived structures may not necessarily be the product of just the individual inter-point distances, but may be influenced by other aspects of the display such as symmetries, regularities and nearest-neighbor distances (e.g., Van der Helm, 1994). In general, the role of the interaction of visual attributes in contributing to the perception of structure in data visualization has been previously noted (Dastani, 2002; but see also Garner, 1974). Some research has also investigated the perception of structures in dot arrays like those in the 2D displays in our experiment (Compton and

[2]Tenenbaum et al. (2002) present a technique for determining radius size for the fixed radius version of ISOMAP (the ε-graph method) that helps preserve the correct topology. This was published after our experiment had commenced.

Table 5
Mauchly's test of sphericity

Within-subjects effect	Dependent measure	Mauchly's W (df)	Significance	Epsilon (greenhouse-geisser)
Question	Accuracy	0.468 (9)	<0.001**	0.776
	Docs accessed	0.710 (9)	0.002**	0.862
	Confidence	0.706 (9)	0.002**	0.870
	Response time	0.363 (9)	<0.001**	0.664
	NN moves	0.561 (14)	<0.001**	0.769

Note: **$p<0.01$.

Logan, 1993; Vickers et al., submitted). However, the specific question of the accuracy of perceived structures in data visualization displays has not been addressed. This issue is an important theoretical one that has the potential to benefit both the design and evaluation of display algorithms.

Finally, one of the limitations in generalizing from the results of this particular study is that the findings may be linked to the type of documents used. Published articles like the ABS NewsMail texts in our experiment are professionally edited, highly structured, and well written. In contrast, much of the data that might be visualized in the real world will be less structured. In order to determine the broader usefulness of such data visualization techniques it will be necessary to test their effectiveness with a variety of document sets.

Acknowledgments

The research reported in this paper was supported by the Australian Defence Science and Technology Organisation. We wish to thank Joanne Spadavecchia and Robyn Vast for conducting the experiment, Lama Chandrasena for creating the visualization interface, Kym Preiss for assistance in the spatial analysis of the arrays and Louise Mullen, Brandon Pincombe, Ian Coat, Michael Webb, Simon Dennis, Daniel Navarro, Kathryn Parsons and Agata McCormac and several anonymous reviewers for their assistance and helpful suggestions.

Appendix A

A starting document is selected and the second document in the Ordered List is that with the highest similarity rating with the first document (excluding self-similarity). The third document is that with the highest similarity rating with the second document from the remaining documents that have yet to be added to the list. This process continues until all documents have been selected. In cases where there are two or more pairwise similarities with equal highest values, the document added to the list is randomly selected from these documents. In addition, this algorithm was repeated on all possible starting documents and the final solution selected was that which had the highest sum of pairwise similarities between neighboring documents in the list.

Appendix B

Both algorithms were optimized with respect to the *Variance Accounted For* (VAF) in the solution in comparison to the empirical space.[3] The MDS algorithm was tested on 100 iterations while the ISOMAP algorithm was tested across both versions—for the K-NN variant, all valid values of K were tested while for the fixed radius form, values of ε were sampled at regular intervals from within the upper and lower bounds of ε that provided valid solutions. It should be noted, however, that the VAF measures from MDS and ISOMAP solutions are not equivalent because the proximity matrix fit by the two different techniques is likely to be different. Therefore no attempt was made to compare performance between the two algorithms on this measure. In this sense, the visualizations represent the best that could be achieved in a practical application using MDS and ISOMAP methods.

Appendix C

See Table 5 for details.

References

Allan, J., Leuski, A., Swan, R., Byrd, D., 2001. Evaluating combinations of ranked lists and visualizations of inter-document similarity. Information processing and Management 37, 435–458.

Basalaj, W., 2000. Proximity visualisation of abstract data. Unpublished doctoral dissertation, University of Cambridge Computer Laboratory.

Balasubramanian, M., Schwartz, E.L., 2002. The ISOMAP algorithm and topological stability. Science 7 (a), 295.

Butavicius, M.A., Lee, M.D., Spadavecchia, J.C., Vast, R.L., 2003. An empirical evaluation of four data visualization techniques for displaying document corpora. Defence Human Factors Special Interest Group Meeting, 12–14 November 2003, DSTO Canberra, Australia.

Chen, H., Dumais, S.T., 2000. Bringing order to the web: automatically categorising search results. In: Turner, T., Szwillus, G. (Eds.), Proceedings of CHI '00, Conference on Human Factors in Computing Systems, The Hague, The Netherlands. ACM, New York, pp. 145–152.

[3] It should be noted that there are more sophisticated measures of fit such as Normalized Stress and Kruskal's stress that can take into account scale variation between the outputs of the different layout algorithms. Such measures would be more appropriate in comparing MDS and ISOMAP solutions than the VAF (Basalaj, 2000). Thankyou to an anonymous reviewer for pointing this out.

Compton, B.J., Logan, G.D., 1993. Evaluating a computational model of perceptual grouping by proximity. Perception & Psychophysics 53 (4), 403–421.

Cox, T.F., Cox, M.A.A., 1994. Multidimensional Scaling. Chapman & Hall, London.

Dastani, M., 2002. The role of visual perception in data visualization. Journal of Visual Languages and Computing 13 (6), 601–622.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnass, G.W., Harshman, R.A., 1990. Indexing by latent semantic analysis. Journal of the American Society of Information Science 41 (6), 391–407.

Diggle, P.J., 1983. Statistical Analysis of Spatial Point Patterns. Academic Press, New York.

Donoho, D.L., Grimes, C., submitted. Local ISOMAP perfectly recovers the underlying parametrization of occluded/lacunary libraries of articulated images. IEEE Computer Vision and Pattern Recognition.

Fechner, G.T., 1966. Elements of Psychophysics. Holt, Rinehart and Winston, New York (Trans. H.E. Adler).

Frøkjær, E., Hertzums, M., Hornbæk, K., 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: Turner, T., Szwillus, G. (Eds.), Proceedings of CHI '00, Conference on Human Factors in Computing Systems, The Hague, The Netherlands. ACM, New York.

Garner, W.R., 1974. The Processing of Information and Structure. Lawrence Erlbaum Associates, Potomac, MD.

Graham, M., Kennedy, J., Benyon, D., 2000. Towards a methodology for developing visualizations. International Journal of Human–Computer Studies 53 (5), 789–807.

Kruschke, J.K., 1992. ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review 99 (1), 22–44.

Lee, M.D., 1999. Algorithms for representing similarity data. DSTO Research Report, DSTO-RR-0152.

Lee, M.D., Vickers, D., 1998. Psychological approaches to data visualization. DSTO Research Report, DSTO-RR-0135.

Lee, M.D., Butavicius, M.A., Reilly, R.E., 2003a. Visualizations of binary data: a comparative evaluation. International Journal of Human–Computer Studies 59, 569–602.

Lee, M.D., Reilly, R.E., Butavicius, M.A., 2003b. An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data. In: Pattison, T., Thomas, B. (Eds.), Proceedings of the Australian Symposium on Information Visualization (inVis.au'03), Adelaide, Australia. Australian Computer Society press, Darlinghurst, Australia, pp. 1–10.

Lee, M.D., Pincombe, B.M., Welsh, M.B., 2005. An empirical evaluation of models of text document similarity. In: Bara, B.G., Barsalou, L., Bucciarelli, M. (Eds.), Proceedings of the 27th Annual Meeting of the Cognitive Science Society, Stresa, Italy. Cognitive Science Society, Austin, TX, pp. 1254–1259.

Levin, C.A., Haber, R.N., 1993. Visual angle as a determinant of perceived interobject distance. Perception and Psychophysics 54 (2), 250–259.

Leuski, A., Allen, J., 2000. Improving interactive retrieval by combining ranked lists and clustering. In: Mariani, J., Harman, D. (Eds.), Proceedings of RIAO'2000, Paris. ACM, New York, pp. 665–681.

Lowe, D., Tipping, M.E., 1996. Feed-forward neural networks and topographic mapping for exploratory data analysis. Neural Computing and Applications 4, 83–95.

Mao, J., Jain, A.K., 1995. Artificial neural networks for feature extraction and multivariate data projection. IEEE Transactions on Neural Networks 6 (2), 296–317.

Morse, E., Lewis, M., 1997. Why information visualizations sometimes fail. In: Proceedings of IEEE international Conference on Systems Man and Cybernetics, Orlando, FL. IEEE Press, Los Alamitos, CA, pp. 1680–1685.

Morse, E., Lewis, M., Olsen, K.A., 2002. Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical, and "spring" displays. Journal of the American Society for Information Science 53 (1), 28–40.

Navarro, D.J., Lee, M.D., 2001. Spatial visualization of document similarity. In: Copas, C. (Ed.), Proceedings of the Defence Human Factors Special Interest Group Meeting, Adelaide, Australia. DSTO, Edinburgh, SA, pp. 39–44.

Nosofsky, R.M., 1986. Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General 115 (1), 39–57.

Preiss, K., 2006. A theoretical and computational investigation into aspects of human visual perception: proximity and transformations in pattern detection and discrimination. Unpublished Doctoral Dissertation, University of Adelaide, Australia. Retrieved March 7, 2007, from 〈http://thesis.library.adelaide.edu.au./public/adt-SUA20060727.135549〉.

Purchase, H.C., 1998. Effects of graph layout. In: Thomas, B., Calder, P. (Eds.), Proceedings of OZCHI 98, Adelaide Australia. IEEE Computer Society, Los Alamitos, CA, pp. 80–86.

Rorvig, M., 1999a. A visual exploration of the orderliness of TREC relevance judgments. Journal of the American Society for Information Science 50 (8), 652–660.

Rorvig, M., 1999b. Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. Journal of the American Society for Information Science 50 (8), 639–651.

Sebrechts, M.M., Cugini, J.V., Vasilakis, J., Miller, M.S., Laskowski, S.J., 1999. Visualization of search results: a comparative evaluation of text, 2D, and 3D interfaces. In: Gey, F., Hearst, M., Tong, R. (Eds.), Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia. ACM Press, New York, pp. 173–181.

Shepard, R.N., 1957. Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika 22 (4), 325–345.

Shepard, R.N., 1987. Toward a universal law of generalization for psychological science. Science 237, 1317–1323.

Shepard, R.N., 1994. Perceptual-cognitive universals as reflections of the world. Psychonomic Bulletin and Review 1 (1), 2–28.

Shneiderman, B., 1998. Designing the User Interface: Strategies for Effective Human-Computer Interaction, third ed. Addison Wesley, Reading, WA.

Smith, A.E., 2000. Machine mapping of document collections: the leximancer system. In: Proceedings of the Fifth Australasian Document Computing Symposium, Sunshine Coast, Australia, DSTC, Sunshine Coast, Australia.

Sutcliffe, A.G., Ennis, M., Hu, J., 2000. Evaluating the effectiveness of visual user interfaces for information retrieval. International Journal of Human–Computer Studies 53, 741–763.

Swan, R.C., Allan, J., 1998. Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In: Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia. ACM Press, NY, pp. 315–323.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for non-linear dimensionality reduction. Science 290, 2319–2323.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2002. Reply to Balasubramanian and Schwartz: The ISOMAP algorithm and topological stability. Science 295, 7a.

Tractinsky, N., Katz, A.S., Ikar, D., 2000. What is beautiful is useful. Interacting with Computers 13 (2), 127–145.

Tversky, A., Hutchinson, J.W., 1986. Nearest neighbor analysis of psychological spaces. Psychological Review 93 (1), 3–22.

Van der Helm, P., 1994. The dynamics of prägnanz. Psychological Research 56, 224–236.

Vickers, D., Preiss, K., Hughes, P., submitted. The role of nearest neighbors in the perception of structure and motion in dot patterns. Perception.

Ware, C., 2000. Information Visualization: Design for Perception. Morgan Kauffman, San Mateo, CA.

Westerman, S.J., Cribbin, T., 2000. Mapping semantic information in virtual space: dimensions, variance, and individual differences. International Journal of Human–Computer Studies 53, 765–787.

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

Wise, J.A., 1999. The ecological approach to text visualization. Journal of the American Society for Information Science 50 (13), 1224–1233.

Wu, M., Fuller, M., Wilkinson, R., 2001. Using clustering and classification approaches in interactive retrieval. Information Processing and Management 37 (3), 459–484.

Zipf, G.K., 1949. Human Behaviour and the Principle of Least Effort. Addison Wesley, Cambridge, MA.