(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Int. J. Human-Computer Studies 70 (2012) 432–449

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

An assessment of email and spontaneous dialog visualizations

Marcus A. Butavicius[a,*], Michael D. Lee[b], Brandon M. Pincombe[a], Louise G. Mullen[a], Daniel J. Navarro[c], Kathryn M. Parsons[a], Agata McCormac[a]

[a]*Defence Science and Technology Organisation, 203L, DSTO, P.O. Box 1500, Edinburgh, SA 5111, Australia*
[b]*Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100, USA*
[c]*School of Psychology, University of Adelaide, Adelaide, SA 5005, Australia*

Received 2 June 2011; received in revised form 22 December 2011; accepted 3 February 2012
Communicated by P. Mulholland
Available online 3 March 2012

Abstract

Two experiments were conducted examining the effectiveness of visualizations of unstructured texts. The first experiment presented transcriptions of unrehearsed dialog and the second used emails. Both experiments showed an advantage in overall performance for semantically structured two-dimensional (2D) spatialized layouts, such as multidimensional scaling (MDS), over structured and non-structured list displays. The second experiment also demonstrated that this advantage is not simply due to the 2D nature of the display, but the combination of 2D display and the semantic structure underpinning it. Without this structure, performance fell to that of a Random List of documents. The effect of document type in this study and in Butavicius and Lee's (2007) study on visualizations of news articles may be partly described by a change in bias on a speed-accuracy trade-off. At one extreme, users were accurate but slow in answering questions based on the dialog texts while, at the other extreme, users were fast but relatively inaccurate when responding to queries about emails. Similarly, users could respond accurately using the non-structured list interface; however, this was at the cost of very long response times and was associated with a technique whereby participants navigated by clicking on neighboring document representations. Implications of these findings for real-world applications are discussed.

Keywords: Data visualization; Multidimensional scaling; Isomap; Empirical evaluation; Human–computer interaction; Email; Spontaneous dialog

1. Introduction

Document visualizations are graphical representations of a set of text documents. The aim of these visualizations is to convey trends and patterns that would be impossible, or very time consuming, to ascertain based on an examination of the individual documents alone. Visualization tools are particularly beneficial when the number of documents in the set to be analyzed is very large. As White, Muresan and Marchionini (2006) have pointed out, document visualization may be of benefit for exploratory data analysis when (1) the search problem is not well defined, (2) the user is not familiar with the problem domain and (3) when multiple points of view need to be considered in investigating the documents. For these reasons, document visualization tools have gained increasing popularity in not only intelligence gathering for security, defense and law enforcement (e.g., Stasko et al., 2008) but also for detecting trends in the domains of science, politics and public opinion (e.g., Clavier and El Ghaoui, 2008; Mothe et al., 2006; Powell, 2004).

A common approach to document visualization involves proximity-based techniques. A specific example of such an approach is a point-based spatialized display (also known as a spatialization) whereby each document is represented

*Corresponding author at: Defence Science and Technology Organisation, 203L, DSTO, P.O. Box 1500, Edinburgh, SA 5111, Australia. Tel.: +61 8 7389 6097; fax: +61 8 7389 6328.
E-mail addresses:
marcus.butavicius@dsto.defence.gov.au (M.A. Butavicius),
mdlee@uci.edu (M.D. Lee),
brandon.pincombe@dsto.defence.gov.au (B.M. Pincombe),
louise.mullen@dsto.defence.gov.au (L.G. Mullen),
daniel.navarro@adelaide.edu.au (D.J. Navarro),
kathryn.parsons@dsto.defence.gov.au (K.M. Parsons),
agata.mccormac@dsto.defence.gov.au (A. McCormac).

1071-5819/$ - see front matter
doi:10.1016/j.ijhcs.2012.02.002

by an icon and the distance between the icons represents the similarity between the documents (e.g., ACQUAINTANCE and PARENTAGE: Liu et al., 2000; LEXIMANCER: Smith, 2000; TEXT GARDEN DOCUMENT ATLAS: Fortuna et al., 2006). That is, the icons for documents that are determined to be similar are positioned closer to each other in the display. This type of layout is consistent with Montello et al.'s (2003) 'First Law of Cognitive Geography' which states that "people believe closer things to be more similar than distant things" (p. 317). In two relatively large scale ($N=45$ and 48) user studies these authors demonstrated empirical evidence for this principle. The success of these displays possibly lies in the ability of the human visual system to detect easily patterns in the display such as clusters and outliers. As Brusco (2007) has pointed out, the ability to partition such point arrays into clusters is one of many visual combinatorial optimization problems for which the human visual system appears to be very well adapted (see also Vickers et al., 2001).

The reliance of spatialized displays on the capabilities and limitations of the human visual processing system, and the need to provide empirical evidence as to their real-world effectiveness, suggests that it is important to study user behavior. In particular, it seems important to study whether or how visualizations facilitate performance on the information-handling tasks they are designed to support. As a result there is a growing body of studies that have conducted empirical research into the effectiveness of visualizations (e.g., Butavicius and Lee, 2007; Cribben and Chen, 2001; Don et al., 2007; Fabrikant et al., 2004, 2006; Tory and Möller, 2004; Tory et al., 2007; Lee et al., 2003; Sanyal et al., 2009; Ware, 2000; Westerman et al., 2005; Westerman and Cribbin, 2000).

While the notion of visualizations such as spatialized displays may have strong intuitive appeal, the empirical support for a performance advantage over traditional list type displays is not unequivocal. Certainly, there is strong evidence that the performance associated with 2D visualizations is better (or at least no worse) than 3D visualizations (Fabrikant, 2002; Newby, 2002; Sebrechts et al, 1999; Westerman and Cribbin, 2000). In addition, Butavicius and Lee (2007) and Tory et al. (2007) have found better performance with spatialized displays when compared to list and landscape displays. However, Cribben and Chen's (2001) user study found better performance, at least for some tasks, for a list-based display when compared to a spatialized display and two network displays that also contained links between related documents. Similarly, Hornbæk and Frøkjær (1999) found that there was no difference in the number of documents retrieved or marked as relevant between text-only search and a visualization display known as a thematic map (essentially a spatialized display with the addition of "theme" words on the display) but that participants took longer when using the thematic map. Qualitative analysis of participants' verbal descriptions of their thoughts and actions in the study suggested that there was a tendency for the users to get 'lost' in the thematic map displays. Finally, Swan and Allan (1998) found only modest improvements in more sophisticated spatialized displays over text-only interfaces. As suggested by Newby (2002), it is still unclear whether "visual interfaces for IR can be more effective than text-based interfaces" (p. 50).

There could be two reasons for the variation in findings regarding list versus spatialized displays in the literature. Firstly, many of the user studies reported in the literature are based on small sample sizes. For example, Swan and Allan (1998), Cribben and Chen (2001) and Hornbæk and Frøkjær (1999) used studies of size 24, 15 and 6 respectively. While such smaller scale studies can be very informative, their results are harder to generalize and lack statistical power. Secondly, there is a great deal of variation between studies in the tools being tested, with many studies employing more complete visualization tools with different combinations of functionalities (e.g., INFOSKY: Granitzer et al., 2004; YAVI: Newby, 2002). As a result, the potential influence of a range of interface and functionality variables is not well controlled across the different experiments.

In this paper, we seek to address the first issue using comparatively large sample sizes ($N=48$ and 49) for our two experiments, and also by using a repeated-measures design to increase statistical power. We attempt to address the second issue by testing specific components of a visualization interface as opposed to a complete visualization tool. A final feature of our approach is that we use human document similarity judgments, rather than machine substitutes, in the construction of the displays. This is particularly important given the inconsistencies between human and machine judgments demonstrated in Lee et al. (2005). The use of human similarities allows us to focus on testing the visual component of the visualization in isolation from the quality of the underlying document similarities (for further discussion see Butavicius and Lee, 2007).

Our approach is similar to 'de-featured' systems (Morse and Lewis, 1997) and BASSTEP methodologies (Morse et al., 2002), in which only basic features are tested or introduced at each stage of the development process. Our approach differs from these methodologies in that we are using a controlled experimental framework for our testing. As Walenstein (2002) has pointed out, testing of the isolated components of software is necessary to comprehend "the abstract principles of the support provided by the tools rather than the interfering details of the particular prototype" (p. 39). This means the findings of this paper may inform the design of a wide range of visualization and information retrieval tools that employ spatialized and list-based displays.

Our study builds on one reported by Butavicius and Lee (2007), who evaluated the performance of 80 participants in an experiment using four different visualization techniques applied to news articles. The displays were a Random

List, an Ordered List and two two-dimensional visualizations using the multidimensional scaling (MDS: Shepard, 1980) and Isomap (Tenenbaum et al., 2000) layout algorithms. All but the Random List display were constructed using human judgments of document similarity from Lee et al. (2005) to ensure that they were structured using a cognitive model of the document space. In the Butavicius and Lee (2007) study, participants performed best – in the sense that they were faster and accessed fewer documents – when using the structured displays and the two-dimensional (2D) spatialized displays outperformed the one-dimensional (1D) lists.

Our study extends Butavicius and Lee's (2007) paper in two main ways. Firstly, in the previous experiment, all the experimental conditions with a 2D layout were structured using algorithms operating on human judgements of document similarity. Therefore, it is not possible to rule out the hypothesis that the performance achieved in these conditions was simply due to the fact that the documents were laid out in the 2D plane. Westerman and Cribbin (2000) showed that increasing the semantic variance accounted for by 2D solutions in spatialized displays (to the order of 50%, 70% and 90%) was found to improve performance on a search task. However, it is not possible to determine from either Butavicius and Lee (2007) or Westerman and Cribbin's (2000) studies whether simply laying document representations out randomly in a 2D plane, without any structuring according to semantic information, may still be advantageous to the user (or conversely, whether such visualizations are indeed worse than unstructured list-based displays). For example, a random 2D spatialized display may allow a user to remember where documents are better than an unordered list of documents. To address this issue, we include a random 2D spatialized display condition in the second experiment in this paper. In so doing, we also address another issue in real world applications of visualization tools, concerned with how a visualization of document space will perform in cases where there is little semantic structure to be found (i.e., the documents are all from disparate topics). Many intelligence and exploratory applications of visualization tools, where the corpora of documents changes frequently, result in the semantic structuring of the space changing rapidly and unpredictably. In addition, distinct semantic structure may be less apparent in visualizations of email and spontaneous speech because, as we discuss shortly, the topicality in such texts can be varied both between and within the documents. It is therefore useful to determine whether visualization tools provide any advantage or disadvantage over conventional list-based displays in these 'worst-case' scenarios.

The second way in which the current paper builds on Butavicius and Lee (2007) is by examining email and transcriptions of telephone conversations. As with most user studies in visualizations, Butavicius and Lee (2007) used well-edited documents. Many previous assessments of visualizations have used similar documents in the form of news articles (e.g., Cribben and Chen, 2001; Granitzer et al., 2004; Experiment II: Newby, 2002) and journal articles (e.g., Hornbæk and Frøkjær, 2003). These sorts of articles are also used extensively to test information retrieval tools in benchmark tests and competitions (Voorhees and Harman, 2005). However, it remains to be seen how well visualization techniques perform when faced with more spontaneous, less polished texts such as unrehearsed conversations and emails.

Spontaneous speech and email are similar to a range of newer communication media involving computer-to-computer interactions including web logs (colloquially known as "blogs"), Internet forums and instant messaging. These fora are increasing in popularity and represent a wealth of information that lends itself to exploration using visualization tools. All of these differ from professionally edited news articles in a number of ways including:

a. *Linguistic features*: particularly in spoken dialog, the presence of features such as speech repairs (Levelt, 1983) and discourse markers (Shiffrin, 1987) can make interpretation of such language difficult (Heeman and Allen, 1997).
b. *Vocabulary*: more conversational or informal communications often feature the use of slang and more fluid language use including specialized vocabulary, emoticons, acronyms and abbreviations.
c. *Information density*: these documents are often characterized by their "loose, unstructured, garrulous or unedited quality" and may be considered to be 'information poor' in comparison to documents that are engineered by communication experts (Toffler, 1970, p. 155). In contrast, engineered documents such as articles, scripts, formal speeches are "highly purposive … [and] pre-processed to eliminate unnecessary repetition" (Toffler, 1970, p. 155).
d. *Breadth of topicality*: rather than being focused on a particular topic, these less-structured documents can cover a range of different topics.

These characteristics can make such communications difficult to analyze for both humans (Hornbæk and Frøkjær, 2003; Ratté et al., 2007) and computers.

Given the dialogic character of these media, many tools for visualizing such archives have centered around presenting and analyzing patterns in the metadata, e.g., the sender, recipient and time/date stamp information associated with an email (e.g., MAILVIEW: Frau et al., 2005) or the author information and thread in newsgroups and web forums (e.g., CONVERSATIONAL LANDSCAPE and LOOM: Donath et al., 1999). Subjective assessments of such visualizations, when used to display hierarchical, correlational and temporal patterns in email archives, have been favorable (Perer et al., 2006; Perer and Smith, 2006). There have also been efforts to represent the content of such communications as well. There are tools documented

in the literature containing spatialized displays for visualizing the entities (i.e., people, places dates and organizations) within an email corpus (JIGSAW: Görg and Stasko, 2008), author's mood (e.g., CONVERSATIONAL LANDSCAPE and LOOM: Donath et al., 1999) as well as content similarities between individual messages via spatializations for blogs (INSPIRE: Gregory et al., 2007; VIZBLOG: Pérez-Quiñones et al., 2007) and topicality clustering of emails (BUZZTRACK: Cselle et al., 2007). However, despite this increased interest to date we are not aware of any empirical study that has tested the performance of visualizations of such communications.

In this paper, we present two experiments that examine how well several proximity-based visualization techniques assist a user in the analysis of spontaneous speech transcripts and email texts. In these displays, we are interested in representing the content of a collection of texts across a number of individuals. This type of display could be of use in a task of an analyst who, for the purposes of business, political or security intelligence gathering, is exploring a corpus of unstructured, spontaneous texts from multiple authors to understand the content of the communications. This type of exploratory analysis of data and documents can play an important role in the work of an intelligence analyst (Gersh et al., 2006; Pirolli and Card, 2005).

2. *Experiment I*: spontaneous speech

In the first experiment, we compared visualization performance using transcriptions of unrehearsed telephone conversations. The types of visualization techniques were similar to those used in Butavicius and Lee (2007) including a Random List, a structured list, a 2D display based on the Isomap algorithm (Tenenbaum et al., 2000) and another 2D display based on multidimensional scaling (MDS: Shepard, 1980). However, as mentioned above, we used transcriptions of unrehearsed, telephone conversations rather than structured, well-edited news articles. We also used a within-subjects design, as opposed to the between-subjects design used by Butavicius and Lee (2007), to improve the statistical power of the design.

2.1. Method

2.1.1. Participants

Forty-eight participants were recruited for the experiment, the majority of whom were students and staff from the University of Adelaide. The average age was 25 years (SD=8) and 26 of the participants were female. Participants received a food voucher redeemable at the University cafeteria to the value of $10 for taking part in the study, except for first year psychology students, who instead received partial course credit. Participants were recruited via bulk email and posters displayed around the University campus.

2.1.2. Documents

Four document sets, consisting of 40 documents each, were selected for use in the experiment. The documents were excerpts from professional transcriptions of natural language taken from the Linguistic Data consortium known as SWITCHBOARD-1 (Godfrey and Holliman, 1997). This set consists of transcripts of telephone conversations in which two participants, who were previously unknown to each other, talked about a prescribed topic. These dialogs were spontaneous not scripted.

All documents were processed to remove notation indicating non-linguistic utterances and sounds. Document excerpts were selected to conform to a hierarchical taxonomy. Specifically, the topics were arranged into five categories (Sports, Crime and Law, Cars, Politics and Miscellaneous), each of which contained a number of topics, as shown in Fig. 1. In the first four categories the documents were all semantically related to documents within the same category. In the Miscellaneous category the documents were such that none of them were judged by the authors to be closely semantically related to any other document in the entire set. In choosing documents in this way, the degree of relatedness between the documents was as consistent as possible across the four sets. This was done to ensure that the document sets were broadly comparable across the different test conditions as well as to assist in constructing information retrieval questions that were also comparable in task and difficulty.

An example of one of the documents from the Crime topic is:

> A: And, seems like all big cities have plenty of that nowadays, doesn't it?
> B: Well, I, that's, sure. I think its statistics, obviously, vary greatly. I always thought of Dallas as being a fairly safe place.
> A: Well, it is, but our crimes up here, as I think it must be in most cities now, but, I was listening to the news the other day and they said they thought a lot of it, the reason it was up so was because of the, so many people are without work nowadays, economy's so bad.
> B: Do you really believe that? I mean, it's been up every year for many years and the economy hasn't been, this bad for so long, has it?
> A: That's a good point. That's just what they quoted over the news.

2.1.3. Questions

Six multiple choice questions, each with four options, were constructed for each document set. The questions all related to factual information (e.g., dates, times, names of places and people), which were specifically chosen so that they would be unlikely to be already known by the participants, but did not require high-level interpretation of the document. Importantly, the questions also indicated

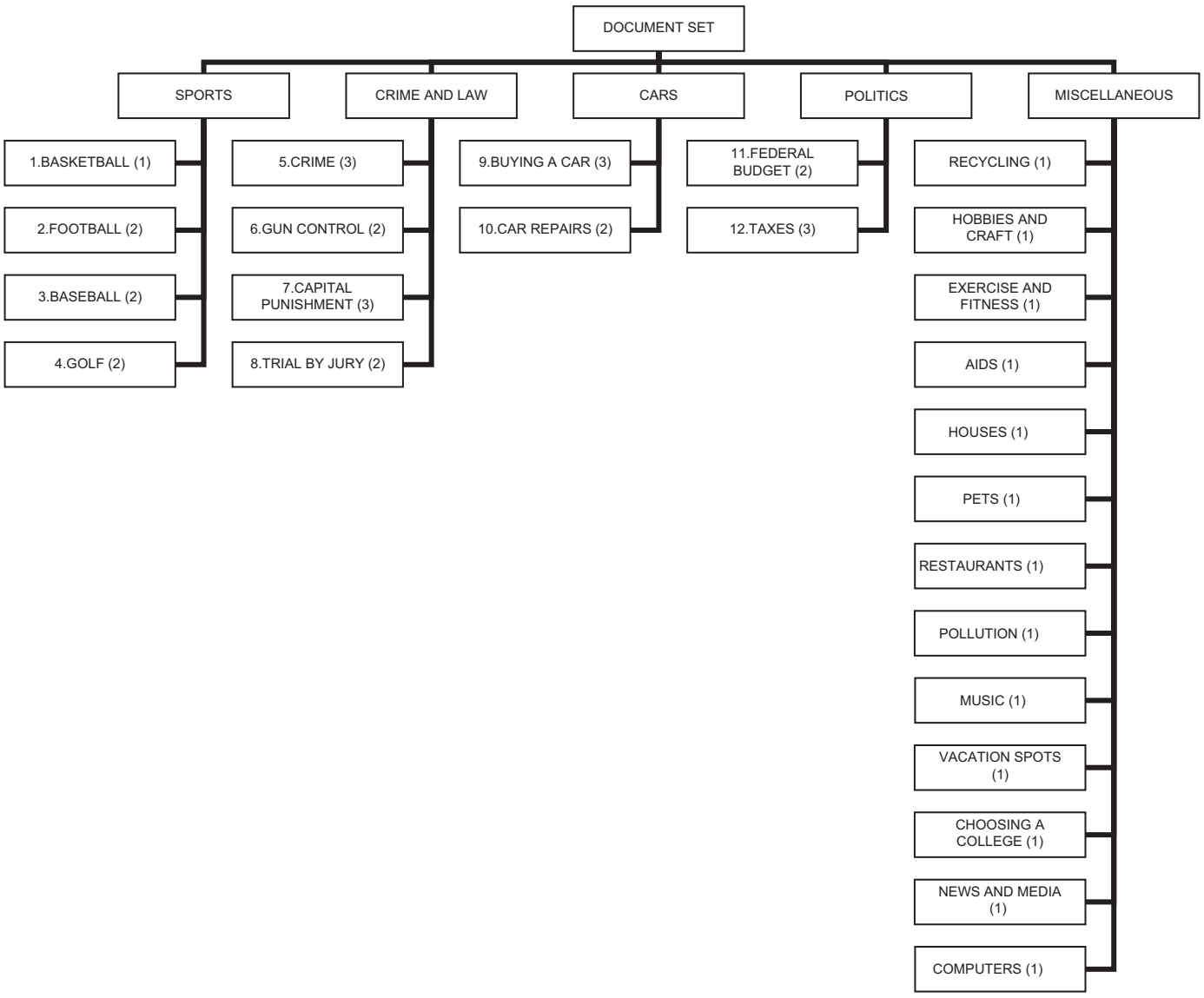

Fig. 1. The arrangement of topics into categories. The number preceding all topic names from the semantically coherent categories indicate the graph labels for Figs. 2 and 3. The number of documents for each topic is shown in brackets. This structure was identical for all four test document sets.

the topic to which the required document related (e.g., a question about baseball statistics explicitly mentioned baseball). The intention of the questions was not to mimic an exact task of an analyst. Given our component-based approach (see also Morse and Lewis, 1997; Morse et al., 2002), we were not testing a complete visualization tool but we were only assessing how well the display layout was consistent with users' expectations. That is, we were testing whether the users were able to understand and make use of the content similarity of the documents. Using the content information as a guide, users can navigate through the display until they find the required document containing the factual information to answer the question. A display that presents the information in a way that is consistent with their expectations of content similarity should be easier for a user to navigate on this type of task. We considered this functional approach to be a better way to assess the effectiveness of the layout than simply relying on subjective judgements of the displays, which are unreliable indicators of effectiveness particularly in the evaluation of interfaces (Frøkjær et al., 2000; Wu et al., 2001; Lee et al., 2003).

Different types of questions were included to allow examination of whether the visualization techniques benefited specific types of document search tasks. The questions varied according to whether one or two documents needed to be retrieved to answer the question. They also varied as to the semantic relationships between the required document(s) and the other documents in the set. More specifically, the question types required access to either:

A. One document outside taxonomy (i.e., from the Miscellaneous category)
B. One document inside taxonomy

C. Two documents both outside taxonomy
D. Two documents from same topic
E. Two documents from same category but different topic
F. Two documents from different category but both still in taxonomy

Each question type related to general types of document comparisons that could be of assistance to the user. For example, finding answers to question type A (one document outside taxonomy) related to searching for outlier documents in the semantic space. Similarly, finding answers to question type D was akin to searching for highly semantically similar documents which should be in close physical proximity in the structured visualizations (e.g., two documents on baseball). Question type E related to the task of finding documents that are semantically related but in a weaker fashion. That is, they were from a broader categorization of documents (e.g., two documents each on different sports). An example of a question from question type D in which participants were required to find two documents from the same topic (Baseball) in the same category (Sports) was:

> With whom are the Rangers baseball team currently negotiating a contract (A) and who is regarded as the player who is their "biggest point of interest" (B)?

A. (A) Rafael Palmeiro and (B) Nolan Ryan
B. (A) Rafael Palmeiro and (B) Kevin Brown
C. (A) Ruben Sierra and (B) Kevin Brown
D. (A) Ruben Sierra and (B) Nolan Ryan

In this example, the correct response was the second option.

2.1.4. Visualizations

Each participant attempted the same questions for each document set, however the document set could be visualized in four different ways. The first was a Random List condition where the documents were arranged randomly in a list. This represents common list-based interfaces where there is no attempt to order according to document similarity.

The remaining three conditions were structured using similarity judgments acquired using a similar approach to Lee et al. (2005). Pairwise similarity judgments were initially gathered from two participants using a computer program that presented every pair of unique documents. These pairs were rated on a five-point scale where one represented "highly unrelated" and five represented "highly related". Using judgments based on averaging across individuals, when significant individual differences exist, could result in a display that does not portray a valid cognitive representation of the document space (Ashby et al., 1994; Lee and Pope, 2003). We examined the differences between the judgements directly. In addition, a third participant provided additional ratings for judgments where the initial two participants differed by two or more points on the similarity scale. These additional judgments served as another means to examine differences between the first two judges. We noted systematic differences between the responses of the two participants who provided all pairwise judgments and used the set of judgments from the one participant who demonstrated the greatest variation in assigning similarity scores as this variation provided more rich semantic information.

The second condition was an Ordered List where the list of documents was structured such that, within the ordinal list constraints, more similar documents were placed next to each other in the list. The algorithm used to generate these lists was the greedy nearest-neighbor algorithm outlined in Butavicius and Lee (2007).

The third and fourth conditions displayed 2D representations of the document similarities. As with the Ordered List, the aim was to place more similar documents closer to each other on the screen. Isomap and MDS both find coordinate pairs for the documents in a 2D space such that the distances between these documents approximate the original pattern of distances between the document pairs as given by the human raters. The primary difference between the two algorithms is that while MDS attempts to find a lower dimensional representation (in this case a 2D solution) directly from the original distances, Isomap firstly processes the original distances by constructing a neighborhood graph based on local proximities (for further details see Tenenbaum et al., 2000).[1] While MDS is already a popular tool in visualizations and has been used as a model for mental representation (Shepard, 1957, 1987), Isomap is theoretically better able to handle non-linear structures that may be present in the original document space (Tenenbaum et al., 2000). The MDS display employed the standard multidimensional scaling layout approach and the Euclidean distance metric (Cox and Cox, 1994).

Table 1 shows the list-based solutions for the first document set represented in terms of category and topic memberships. As can be seen, the Ordered List solution placed all of the documents from the semantically coherent categories (i.e., all but the Miscellaneous group) next to each other. In addition, documents from the same topic were adjacent to each other with the exception of the Sports and Cars categories.

For some document sets, the 2D techniques provided distinctly different types of solutions. The MDS solution for the first document set is shown in Fig. 2. This contrasts

[1] In this study, both algorithms were also optimized with respect to the *Normalized Stress* (Basalaj, 2000) of the solution. The *MDS* algorithm was tested on 100 iterations while both versions of the *Isomap* algorithm were tested. For the *K*-nearest neighbor 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. For further discussion on the optimization of these displays for visualization see Butavicius and Lee (2007).

Table 1
List visualizations with respect to category with topic indicated in brackets and Miscellaneous documents indicated by asterisks.

Position	Ordered	Random
1	*	Sports (basketball)
2	Sports (golf)	*
3	Sports (golf)	Sports (baseball)
4	*	Cars (buying a car)
5	Sports (football)	*
6	Sports (baseball)	Cars (car repairs)
7	Sports (basketball)	Crime and law (capital punishment)
8	Sports (football)	Politics (federal budget)
9	Sports (baseball)	Crime and law (crime)
10	*	Politics (taxes)
11	Cars (buying a car)	Crime and Law (capital punishment)
12	Cars (car repairs)	Politics (federal budget)
13	Cars (buying a car)	*
14	Cars (car repairs)	*
15	Cars (buying a car)	Crime and Law (crime)
16	Politics (taxes)	*
17	Politics (taxes)	Cars (car repairs)
18	Politics (taxes)	Cars (buying a car)
19	Politics (federal budget)	*
20	Politics (federal budget)	*
21	Crime and Law (crime)	Sports (football)
22	Crime and Law (crime)	*
23	Crime and Law (crime)	Crime and Law (gun control)
24	Crime and Law (gun control)	Sports (football)
25	Crime and Law (gun control)	Crime and Law (capital punishment)
26	Crime and Law (capital punishment)	*
27	Crime and Law (capital punishment)	Crime and Law (gun control)
28	Crime and Law (capital punishment)	Crime and Law (trial by jury)
29	Crime and Law (trial by jury)	*
30	Crime and Law (trial by jury)	*
31	*	Sports (golf)
32	*	*
33	*	Politics (taxes)
34	*	Sports (golf)
35	*	Sports (baseball)
36	*	Crime and Law (crime)
37	*	Politics (taxes)
38	*	Cars (car repairs)
39	*	Sports (football)
40	*	*

with the Isomap solution in Fig. 3. Both demonstrate clusters of topically related documents but the Isomap solution demonstrates a distinctive arrangement of these clusters. In the MDS solution, the categories of Sports, Cars and Politics are represented by distinctive clusters although the Crime and Law documents are less consistent with the taxonomy. In the Isomap solution there are approximately four visually distinct clusters—all contain both topically related and non-topically related documents except for one that contains just two non-topically related items. Within these clusters, all the topic groupings are maintained. The subgroups of Cars and Politics are maintained, each in different clusters. Most interesting is the fourth cluster that appears to have organized the similarity between documents contained within it along approximately one dimension. At one end of this dimension are the topics contained in the Crime And Law subgroup and at the other extreme are the Sports documents.

2.1.5. Interface and procedure

These visualizations were presented to participants within a specially designed program. This program not only allowed us to control the presentation of the visualizations and ensure control over which visualization was assigned to which document set, but it also allowed us to log the actions of the participants for later analysis. As we have emphasized throughout, we only used a simplified visualization interface because we were not testing a complete visualization tool but only the effectiveness of the visual component in isolation.

The experiments were all conducted at the University of Adelaide's School of Psychology computing laboratories. At the start of each experimental session, a research

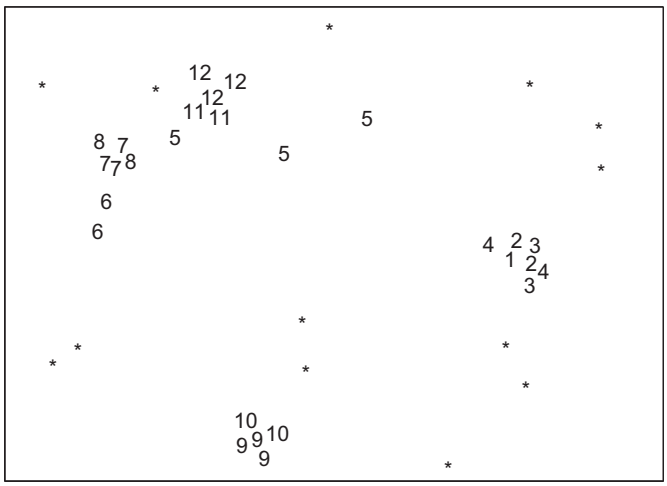

Fig. 2. Representation of the MDS solution for the first document set. The topic membership is indicated by a number for documents belonging to semantically coherent categories and by an asterisk for those in the 'Miscellaneous' category. The graph labels are contained in Fig. 1.

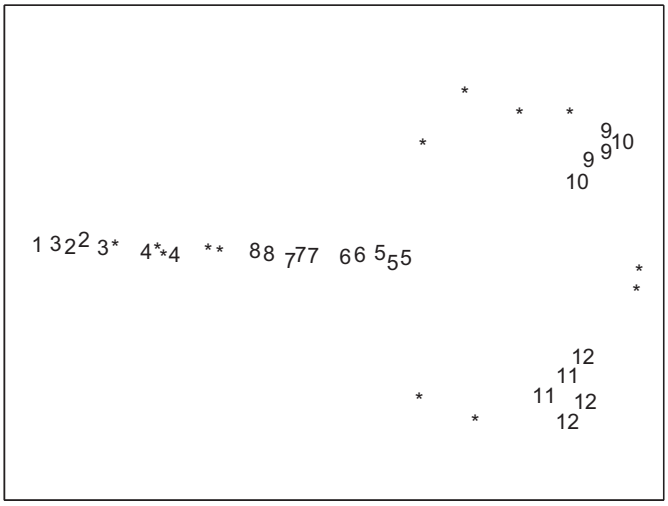

Fig. 3. Representation of the Isomap solution for the first document set. The topic membership is indicated by a number for documents belonging to semantically coherent categories and by an asterisk for those in the 'Miscellaneous' category. The graph labels are contained in Fig. 1.

assistant explained the nature of the experiment. They described the interface and then talked the participant through a practice question in order for the participant to become familiar with the interface and visualization. The practice question was based on documents not present in the main experiment.

Participants were then presented with a series of questions, using each of the four different display types. The design was a modified Latin-square design such that, for each of the two blocks of 24 participants, there were all possible combinations of document sets and visualizations, all possible permutations of visualizations, and there was random assignment of visualization order to visualization-document assignments. This design ensured that there was control for interaction effects between visualizations and document sets as well as order effects associated with mental fatigue or learning effects.

All six questions for each document set were completed for each visualization before moving onto the next visualization. The order of the questions for each document set was randomized before the start of the experiment and these permutations were repeated for all the participants. The order of the multiple-choice answers was randomized for each question. There was no time-limit to complete the questions.

Figs. 4 and 5 show two screenshots of the interface with an MDS 2D visualization and a random-list display,

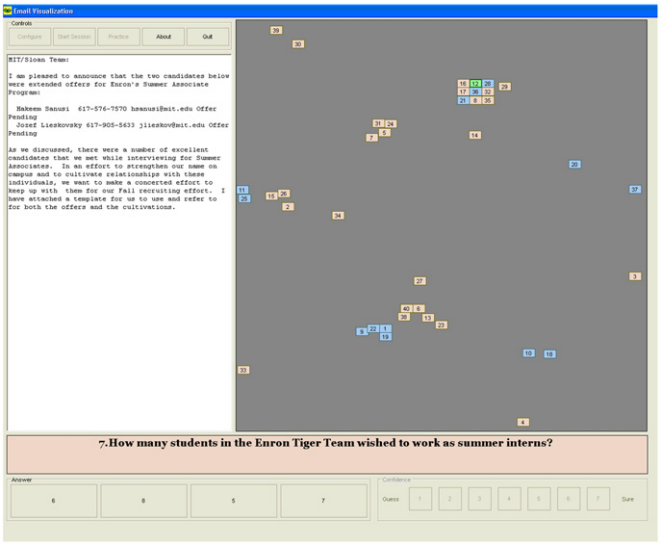

Fig. 4. Experiment interface showing an MDS 2D visualization of one of the Enron email document sets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

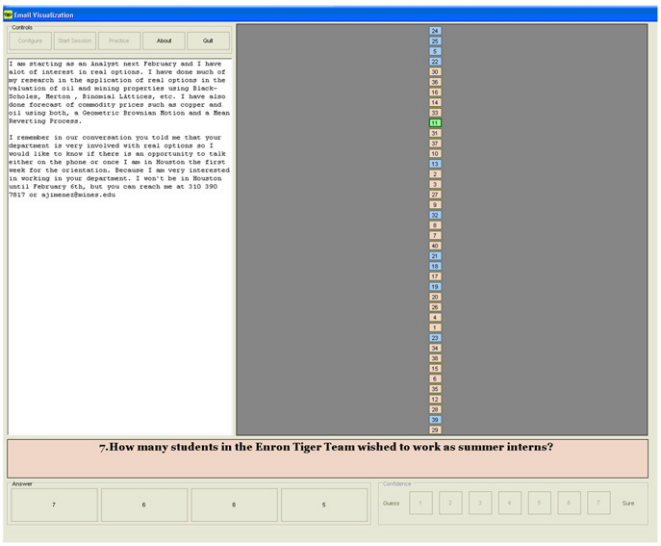

Fig. 5. Experiment interface showing a Random List display of one of the Enron email document sets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

respectively. The top right pane contained the visualization of the corpus. The color of the icons indicated the status of the document representation with respect to the search actions that had been completed for that particular question. Specifically, colors indicated whether the document representation had been accessed (blue) or not (tan) and also which document representation was currently active (green). After the participant had answered the question and provided a confidence rating the color of all the document icons were reset to tan for the following question. The background of the visualization was light gray.

The text of the active document was displayed in the top left pane, as shown in Fig. 4. Directly below the visualization and the document pane was the question pane, colored tan. Underneath this were the response options to the question and confidence ratings on the left and right of the page respectively, both in gray and represented by radio buttons. In order to ensure that participants provided a response option to the question before the confidence ratings, the confidence ratings were inactivated until a response option was selected. After the question had been answered the response options and document icons were inactivated until a confidence response had been selected to ensure that participants could not change their answers.

User actions during an experimental session were logged by the program. This included not only which document icons were clicked on but the answers to the questions and the confidence ratings provided. The log included time stamps for each of these actions to assist with the analysis of timing information.

2.2. Results

In this analysis, the terms 'small', 'medium' and 'large' refer to the magnitude of effect sizes as per Cohen's (1988) guidelines. A four by six way repeated measures analysis of variance (RMANOVA) was performed on the variable of response time with four levels for display (Random List, Ordered List, Isomap, MDS) and six levels for question type.[2] Response time varied significantly between display type (Wilks' Lambda = .803, $F(3,45) = 3.671$, $p = .019$) with a medium effect size (multivariate $\eta_p^2 = .197$). In Fig. 6, there is a trend visible indicating a speed advantage for the 2D structured visualizations, especially over the Random List condition. Bonferroni multiple comparisons indicated a significant difference between the Random List and Isomap displays ($\text{Mean}_{\text{difference}} = 25.61$ s, $\text{CI}_{95\%} = [18.59, 49.34]$, SE = 86.21, $p = .028$) and a difference that was close to significance at the .05 level between Random List and MDS displays ($\text{Mean}_{\text{difference}} = 26.16$ s, $\text{CI}_{95\%} = [-29.9, 52.61]$, SE = 96.05, $p = .054$). In summary, there was an advantage in the 2D structured displays over the Random List condition that amounted to around 25 s on each question. This means that the MDS and Isomap displays allowed users to, on average, answer questions in approximately 83% of the time taken when using the Random List.

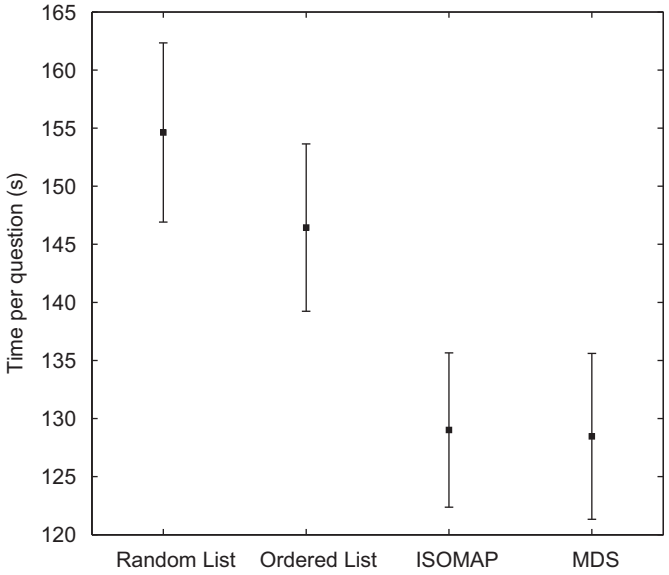

Fig. 6. *Experiment I*: mean response time (s) across the four conditions. One standard error is shown about the mean.

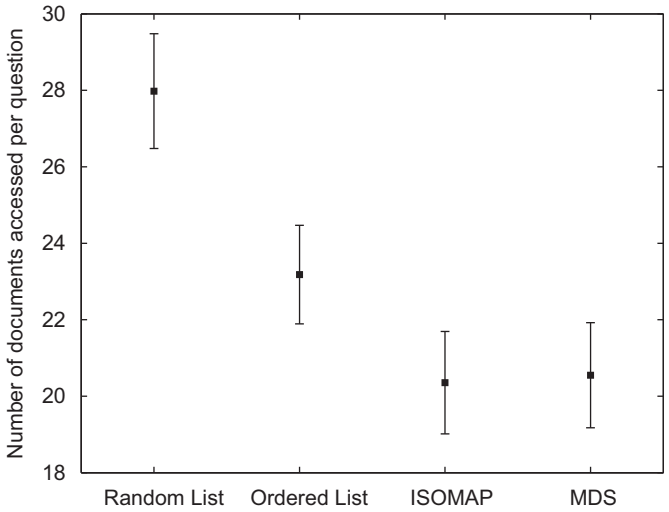

Fig. 7. *Experiment I*: mean number of documents accessed per question across the four conditions. One standard error is shown about the mean.

A similar RMANOVA for the dependent variable of the documents accessed demonstrated an advantage in the structured visualizations over the Random List. The number of documents accessed varied significantly across display type (Wilks' Lambda = .567, $F(3,45) = 11.461$, $p < .001$), with a large effect size (multivariate $\eta_p^2 = .433$). Bonferroni comparisons confirmed the trend visible in Fig. 7 that fewer documents were accessed using the

[2]Because there was evidence of violations of the sphericity assumption in this and other ANOVAs in this paper, we chose to report the more conservative multivariate F values (Wilks' Lambda). For the analyses with time as the dependent variable, although the distributions were not perfectly normal, the deviations from normality were not considered to be severe enough to invalidate the use of standard ANOVA.

structured visualizations compared to Random List ($Mean_{Random\ List–Ordered\ List}=4.799$, $CI_{95\%}=[.771, 8.826]$, $SE=1.462$, $p=.012$; $Mean_{Random\ List–MDS}=7.625$, $CI_{95\%}=[3.844, 11.406]$, $SE=1.372$, $p<.001$; $Mean_{Random\ List–Isomap}=7.431$, $CI_{95\%}=[3.551, 11.311]$, $SE=1.409$, $p<.001$). Although there were fewer documents accessed in the structured 2D displays than the 1D structured display, none of these mean differences were significant and the associated 95% confidence intervals all included zero. In summary, answering questions using structured displays required, on average, accessing 4.8 to 7.6 fewer documents than were needed when a Random List was used. This amounts to a mean reduction in the number of documents accessed of 17% to 27% across the different structured displays.

We examined users search strategies by noting the relative positions of sequentially accessed documents. As outlined in Butavicius and Lee (2007), one way a user may navigate a display is by clicking on nearest-neighbor document representations. When a user has identified a cluster whose topicality is that of a document they are searching for, this is a sensible strategy (i.e., the required document will likely be close to another document of the same topic). However, a heavy reliance on clicking on nearest neighboring document representations may represent a default strategy. If a user cannot perceive, or chooses not to rely on, the semantic structure in a display, navigating the display in this manner represents a brute force technique that minimizes the mouse movements – in a manner consistent with Zipf's (1949) *principle of least effort* – but which still guarantees that the user will eventually find the desired document.

The proportion of sequentially accessed documents that were nearest neighbors (NNs) is displayed in Fig. 8 for each visualization. The proportion of NNs varied

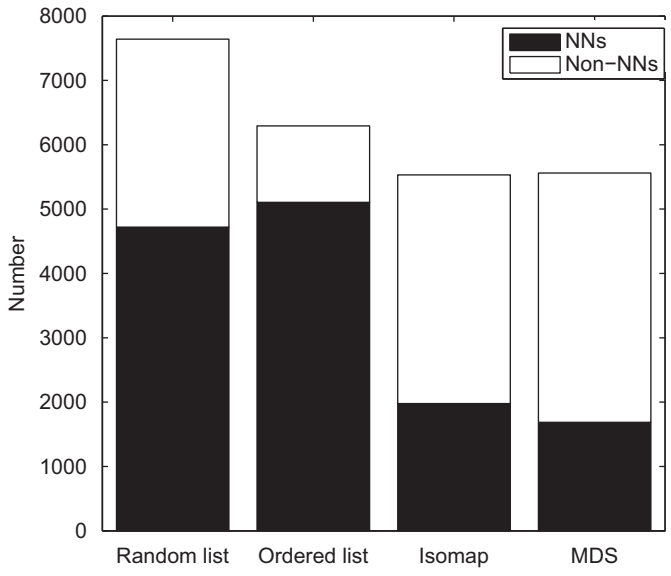

Fig. 8. *Experiment I*: stacked bar graphs of the number of moves made by participants in the display across the four conditions and classified by nearest neighbor (black) and non-nearest neighbor moves (white).

Table 2
Bonferroni comparisons for the proportion of nearest neighbor moves.

Display #1	Display #2	Mean difference$_{\#1-\#2}$ (SE)	p	$CI_{95\%}$
Random List	Ordered List	−.133 (.048)	.045*	[−.264, −.002]
	Isomap	.25 (.051)	<.001**	[.109,.39]
	MDS	.3 (.05)	<.001**	[.163,.438]
Ordered list	Isomap	.383 (.027)	<.001**	[.31,.456]
	MDS	.434 (.029)	<.001**	[.353, 514]
MDS	Isomap	.051 (.016)	.015*	[−.095, −.007]

*Significant at the .05 alpha levels.
**Significant at the .001 alpha levels.

significantly between the displays (Wilks' Lambda = .168, $F(3,45)=74.13$, $p<.001$) with a large effect size ($\eta_p^2=.832$), indicating that over half of the variability in the proportion of nearest neighbor moves was associated with differences between displays. In addition, all of the Bonferroni comparisons (given in Table 2) were significant and those between 1D and 2D displays were significant at $\alpha=.001$.

A similar RMANOVA for confidence responses showed relatively less variation across the different displays (multivariate $\eta_p^2=.18$) although the overall difference was still statistically significant (Wilks' Lambda = .82, $F(3,45)=3.295$, $p=.029$). The only significant Bonferroni comparison was between the Random List and the Ordered List and this latter display was associated with the highest overall average confidence ($Mean_{Random\ List–Ordered\ List}=-.271$, $CI_{95\%}=[-.516, -.026]$, $SE=.089$, $p=.023$). However, the difference only amounted to a half a point difference on a 7 point rating scale. In addition, an examination of the raw data demonstrated that the overall bias of responses was towards highly confident, with 68% of all responses associated with the highest confidence score possible.

The overall accuracy rate was very high at 93%. There was no clear evidence that the type of display influenced how accurately the participants answered the questions. Examination of the graph in Fig. 9 demonstrates no meaningful trend in mean differences, with a substantial overlap in variance across the four display types. The overall RMANOVA on accuracy was not significant (Wilks' Lambda = .898, $F(3,45)=1.703$, $p=.18$), with only a medium effect size (multivariate $\eta_p^2=.102$).

While some question types were more difficult than others, the actual display condition did not change performance differently for different questions. Rather, it influenced performance over all the questions similarly. This is consistent with Butavicius and Lee's (2007) finding that a good visualization can assist a user in various tasks including finding outlier or exceptional documents as well as finding documents that are related to or consistent with other documents in the set. With the exception of the

proportion of nearest neighboring documents selected, overall performance varied significantly across the six different question types, as shown in Table 3. However, there was no evidence that this pattern varied significantly between the displays as demonstrated by the lack of any interaction effect between display and question type.

Correlations were calculated between all of the dependent variables. Spearman's rho (ρ) was used due to the non-normality of some of the response distributions. Initially, all correlations were calculated separately for the different question sets. However, there were no meaningful differences in the trends between the question sets so the results reported here are based on data collapsed across question type. Not surprisingly, the most convincing trend was a large positive correlation between the time taken to respond to a question and the number of documents accessed ($\rho = .799$ [$CI_{95\%}$: .77, .825], $p < .001$, $N = 1152$). This effect was similar across all displays such that 64% of the variation in time taken to respond was associated with the number of documents accessed. Interestingly, there were also overall medium sized effects indicating that an increase in the number of documents accessed was also associated with reduced accuracy ($\rho = -.145$ [$CI_{95\%}$: $-.201, -.088$], $p < .001$, $N = 1152$) and reduced confidence ($\rho = -.141$ [$CI_{95\%}$: $-.197, -.084$], $p < .001$, $N = 1152$).

The second strongest and most consistent trend in terms of effect size, was the correlation between confidence and accuracy ($\rho = .461$ [$CI_{95\%}$:.414, .505], $p < .001$, $N = 1152$). Not surprisingly, this suggests that when participants answered the question correctly they were most confident of their answers. Interestingly, longer response times were associated with a higher proportion of nearest neighbor moves in the two list-based displays ($\rho_{\text{Random List}} = .154$ [$CI_{95\%}$:.039, .265], $p < .001$, $N = 1152$; $\rho_{\text{Ordered List}} = .14$ [$CI_{95\%}$:.025, .252], $p < .001$, $N = 1152$), and this effect was medium sized in both. This correlation is consistent with the idea that the nearest neighbor moves were associated with a default search strategy that is less directed than one based on interpretations of a display's structure.

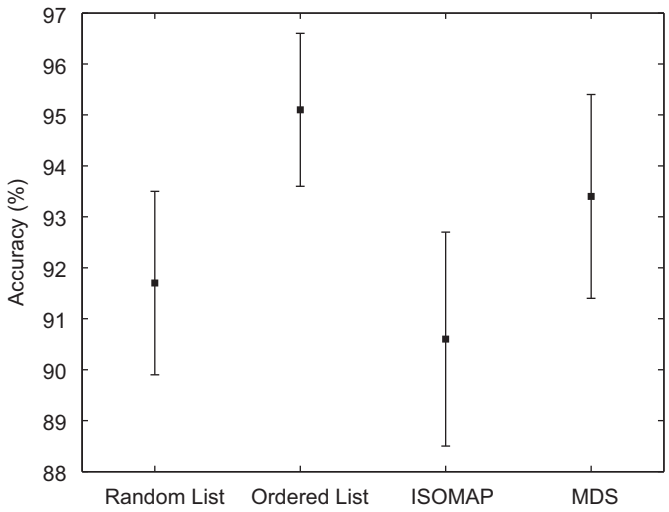

Fig. 9. *Experiment I*: mean accuracy scores across the four conditions. One standard error is shown about the mean.

2.3. Summary

In summary, users performed better with the structured 2D visualizations of the transcriptions of spontaneous speech than the Random List approach. They were 25 s faster and accessed 5–8 fewer documents per question. Proportionally, this amounted to 17% less time and 17–27% fewer documents. Overall, performance on the structured lists was better than the Random Lists but inferior to the 2D structured displays. There were no significant differences in terms of accuracy in performance across the different display types. These results are similar to those of Butavicius and Lee (2007), with the qualification that the performance advantage is expressed in different ways. In particular, in the experiment on news articles the advantages were expressed in terms of accuracy and not speed. Interestingly, while the two 2D visualization approaches in this experiment produced distinctly different interpretations of the semantic structure of the corpora (see Figs. 2 and 3), there was no significant performance difference between them.

Table 3
Repeated measures analysis of variance (effects of question and question by display type).

Source	Measure	Wilk's Lambda	df (error)	F	p	η_p^2
Question	Accuracy	.724	5(43)	3.282	.013*	.276
	Documents accessed	.298	5(43)	20.221	< .001**	.702
	Confidence	.498	5(43)	8.676	< .001**	.502
	Time	.292	5(43)	20.849	< .001**	.708
	Proportion NNs	.844	5(43)	1.587	.184	.156
Question *x* display	Accuracy	.578	15(33)	1.472	.126	.422
	Documents accessed	.543	15(33)	1.854	.069	.457
	Confidence	.734	15(33)	.798	.671	.266
	Time	.555	15(33)	1.765	.085	.445
	Proportion NNs	.633	15(33)	1.275	.271	.367

*Significant at the .05 alpha levels.
**Significant at the .001 alpha levels.

3. *Experiment II*: Enron emails

Experiment II differs from Experiment I in two main ways. Firstly, the document set consists of emails from the Enron Corporation data set rather than transcriptions of spoken dialog. During the legal investigation of the Enron Corporation, the Federal Energy Regulatory Commission released a large collection of actual emails from the corporation, containing over 600,000 messages, from approximately 150 employees (Klimt and Yang, 2004). These emails not only contain messages relevant to the legal proceedings, but also contained other work related and private communications.

Secondly, we changed the types of displays tested. One consideration that was not addressed in Experiment I or Butavicius and Lee (2007) is the degree to which the performance advantage afforded by the structured 2D display is attributable to the fact that the document icons are presented in a 2D plane, and not the cognitive structure that is represented. Westerman and Cribbin (2000) have demonstrated empirically that increasing the degree of semantic information in a 2D visualization improves performance. However, no previous study has examined whether a 2D layout provides any advantage over a 1D layout in the absence of any cognitive or semantic structure. For example, it is conceivable that representing documents in this way helps a user to remember where previously accessed documents are, even if the arrangement of these documents does not reflect semantic similarity.

To address this issue, the second experiment included a random 2D condition to separate the effects of dimensionality and structure on performance. This condition also simulates cases that occur in many real world applications of visualization techniques, where the underlying semantic structure of the corpus is sparse such that there are few natural groupings of documents to be discovered. Such situations may be more frequent when spontaneous language is used. In other words, this experiment is addressing the question of how helpful (or unhelpful) visualizations may be when there is little structure in the information being displayed.

3.1. Method

3.1.1. Participants

Forty-nine participants were recruited for the experiment the majority of whom were students and staff from the University of Adelaide. None of the participants had taken part in the first experiment. The mean age was 27 years (SD=8) and 27 of the participants were female. Participants received a $10 gift certificate from a local multimedia store for taking part in the study, except for first year psychology students who instead received partial course credit.

3.1.2. Documents

The document sets consisted of forty emails each. To help create sets where there were equal numbers of documents on the same topic the *topics model* (Griffiths and Steyvers, 2004) was used to examine and search for emails on similar topics. The results of the topics model analysis were only used for preliminary searches and all documents were ultimately assessed for topicality by one or more of the authors.

The topics were classified into two larger categories—WORK and NON-WORK related emails. In the WORK area, the selected topics (with the number of documents in each set pertaining to that topic indicated in brackets) included:

- *9/11 (3)*—pertaining to the terrorist attacks on the United States and the potential financial effects on the Enron corporation.
- *CPUC (5)*—communications, mostly internal to Enron, regarding the pending investigation into Enron and its dealing with other US energy brokers by the California Public Utilities Commission (CPUC).
- *El Paso (3)*—the dealings of the El Paso Natural Gas Company and particularly the CPUC and Federal Energy Regulatory Commission (FERC) investigation into their anticompetitive conduct.
- *Summer Internships (7)*—the soliciting, hiring or managing of US college students employed by Enron for short term projects over the mid-semester break.
- *Other Recruitment (4)*—any recruitment related correspondence excluding Summer Internships.
- *Outlook (2)*—the impending corporate-wide switch from Lotus Notes to the Microsoft Outlook Email software system.
- *Software (4)*—the installation, upgrade and maintenance of software used within Enron.
- *Training (4)*—the planning, conduct and materials for training courses and seminars organized for Enron employees.

For the NON-WORK area the topics included:

- *Charity Events (2)*—Charity Events organized within Enron primarily for Enron employees.
- *Personal chit-chat (3)*—non-work related correspondence involving at least one Enron employee. Often includes communications with spouses, friends and relatives.
- *Jokes (1)*—emails containing jokes deliberately distributed by/among Enron employees often involving several recipients per message.
- *Non Personal Non-Work (2)*—otherwise known as email spam this consists of unsolicited or undesired bulk email messages received by at least one Enron employee.

Participants were not provided with subject or topic information, and any signature information within the emails was removed. In order to ensure that the emails could be clearly displayed on the interface, the documents

were quite short, with fewer than 500 words each. The document shown below provides an example of the 'Other Recruitment' topic:

> Joe–
> As a follow up on our meeting last week, I'm working with Rick Causey and CAOs for ENA and Enron Europe to identify potential candidates and to refine our job description for the local hire we want to recruit permanently. Rick wants to be closely involved in those decisions. Would you please forward to me some of the handouts you had with you or may have updated by now that address the business environment, Gantt chart/timeline, office scope, timing of business transactions etc. to aid in communicating Tokyo needs? I'm not sure if you sent anything to Sally, but I don't believe I've seen anything yet.
> Thank you,
> Cassandra

3.1.3. Questions

Participants answered seven multiple choice questions for each of the four document sets, meaning that all participants answered 28 questions in total. Each question had four possible choices. As with Experiment I, responses did not require high level analysis of the emails, but the retrieval of clearly stated facts such as dates, times and names within the documents.

The experiment consisted of seven different types of questions that varied in the number of documents that needed to be accessed that contained the required information and the relationship between the document(s) and the rest of the set. For example, some questions required access to only one document, some required access to two documents from the same topic, and some required access to two documents from different topics. The questions also differed according to whether they were WORK or NON-WORK emails. The different question types are shown below:

A. One document from a WORK topic
B. One document from a NON-WORK topic
C. Two documents from the same WORK topic.
D. Two documents from the same NON-WORK topic.
E. Two documents from different WORK topics.
F. Two documents from different NON-WORK topics.
G. Two documents from WORK and NON-WORK topics.

An example of a question from the third document set in which participants were required to find two documents from the same work topic (Summer Internships) is:

> (A) Recruiter Vince Kaminski visited Shmuel several years ago with whom? (B) Who is Samantha Ray now recruiting for?

A. (A) Cantekin Dincerler and (B) EPS
B. (A) Aram Sogomonian and (B) EPS
C. (A) Cantekin Dincerler and (B) EES
D. (A) Aram Sogomonian and (B) EES

In this example, the correct response is the second option (B). The order of the response options was randomised for each trial.

3.1.4. Visualizations, interface and experimental design

The four display conditions were a Random List, Ordered List, Random 2D and MDS 2D display. The structured displays were constructed in the same manner as the first experiment and the similarity judgments on which they were applied were also collected in the same way. An example of the Random 2D visualization of one of the Enron document sets used in this experiment is shown in Fig. 10. The interface was identical to that used in Experiment I. Except for replacement of the Isomap display with a Random 2D display, the experimental design was equivalent to that in Experiment I.

3.2. Results

In total, 1372 questions were answered, and 72% (982) were answered correctly. Interestingly, participants were most accurate when using the Random List display as can be seen in Fig. 11. The RMANOVA indicated that there was significant variation in accuracy associated with the different displays (Wilks' Lambda = .828, $F(3, 46) = 3.176$, $p = .033$, multivariate $\eta_p^2 = .172$). Post-hoc tests using a Bonferroni correction for multiple comparisons yielded only the one significant difference associated with the large drop in performance in the Random 2D display compared to the Random List ($\text{Mean}_{\text{Random 2D–Random List}} = 9.9\%$, $\text{CI}_{95\%} = [.7, 19.1]$, SE = .033, $p = .028$).

However, examination of the response times suggests that, overall, performance on the Random List display was

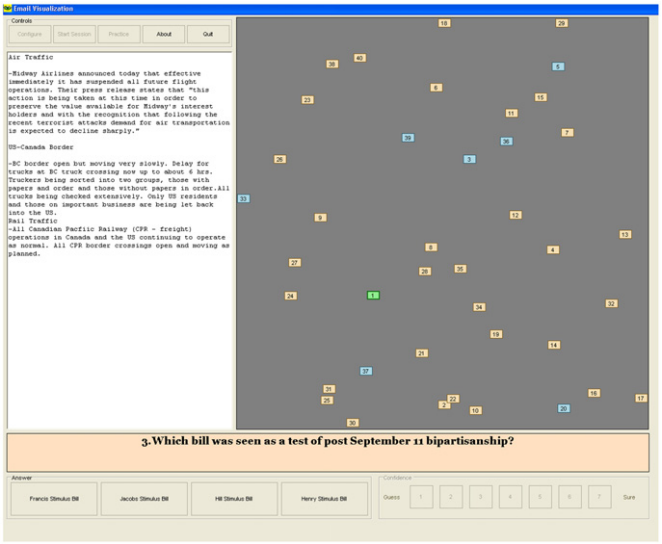

Fig. 10. Experiment interface showing a random 2D visualization of one of the Enron email document sets.

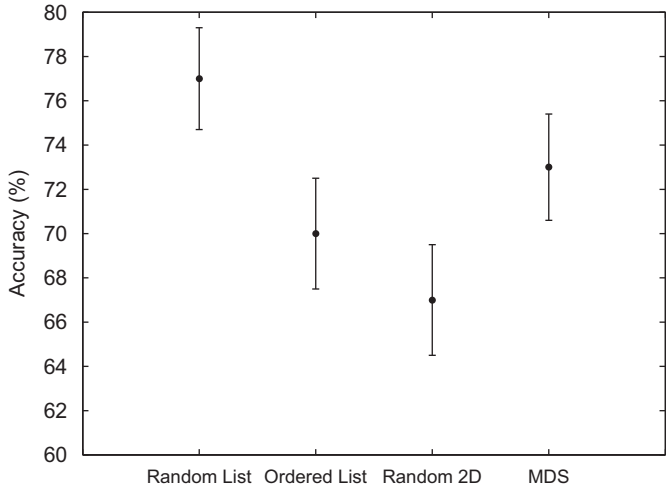

Fig. 11. *Experiment II*: mean accuracy across the four conditions. One standard error is shown about the mean.

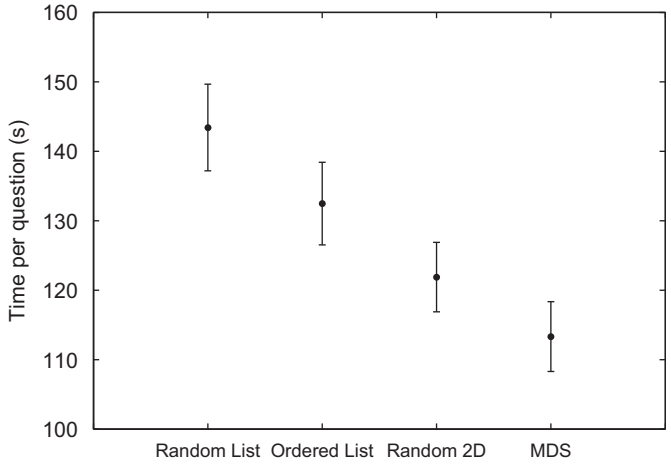

Fig. 12. *Experiment II*: mean response time (s) across the four conditions. One standard error is shown about the mean.

poor and that participants' high accuracy using this technique was due to them trading off speed for accuracy. Response time varied significantly across the visualizations (Wilks' Lambda = .832, $F(3, 46) = 3.106$, $p = .036$, multivariate $\eta_p^2 = .168$) and, as can be seen in Fig. 12, the Random List display took the longest time while the MDS display took the shortest. The Random List display was associated with significantly longer response times than the MDS display ($\text{Mean}_{\text{Random List–MDS}} = 30.11$, $\text{CI}_{95\%} = [.20,60.02]$, SE = 10.87, $p = .048$). This amounts to an average reduction in time taken of 21%. Overall, the mean response time was 127.78 s, with a standard deviation of 103.85 s.

Participants accessed a relatively large proportion of the documents to answer each question, with an average of 25.93 documents (SD = 19.57). Looking at individual trials there was evidence that correct responses were associated with fewer documents accessed, however this trend only accounted for 1% of the variation ($\rho = -.112$ [$CI_{95\%}$: −.164,−.059],

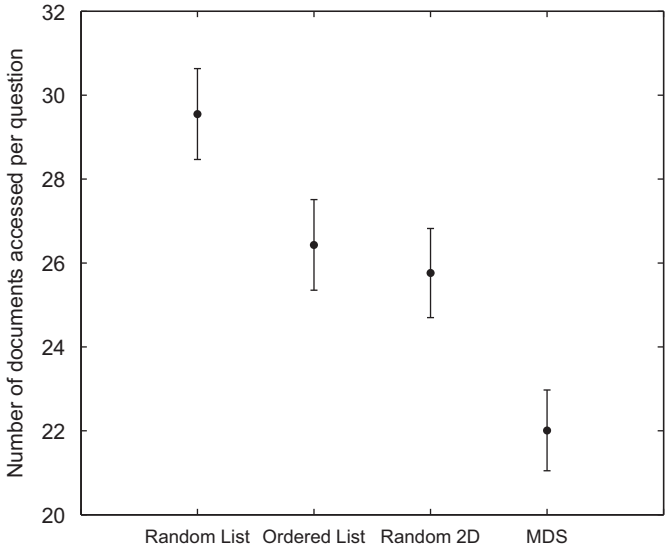

Fig. 13. *Experiment II*: mean number of documents accessed per question across the four conditions. One standard error is shown about the mean.

$p < .001$, $N = 1344$). There was a significant difference in the number of documents accessed between the displays (Wilks' Lambda = .643, $F(3, 46) = 8.522$, $p < .001$, multivariate $\eta_p^2 = .357$). As is visible in Fig. 13, Bonferroni comparisons revealed that MDS was associated with significantly fewer documents accessed than the Random List ($\text{Mean}_{\text{MDS–Random List}} = -7.55$, $\text{CI}_{95\%} = [-11.62,-3.47]$, SE = 1.482, $p < .001$). This amounts to a reduction of 26% in the number of documents accessed compared to the Random List.

Not surprisingly, participants were more confident when the response was correct and far less confident when the response was incorrect. This finding was supported by a strong, positive correlation between confidence and accuracy ($\rho = .61$ [$\text{CI}_{95\%}$:.58,.64], $p < .001$, $N = 1344$). Overall, participants' confidence ratings were quite high, with an average rating of 5.33 (SD = 2.21) and the most common response was the highest confidence rating. There was no significant difference in confidence between displays (Wilks' Lambda = .861, $F(3, 46) = 2.48$, $p = .073$, multivariate $\eta_p^2 = .139$).

There was significant variation between the displays on the proportion of moves that were between NNs (Wilks' Lambda = .072, $F(3, 46) = 197.63$, $p < .001$, multivariate $\eta_p^2 = .928$). As can be seen in Fig. 14, the list displays appeared to attract a higher proportion of NN moves than the 2D displays and this was supported by the Bonferroni comparisons ($\text{Mean}_{\text{Random List–Random 2D}} = .483$, [$\text{CI}_{95\%} = .416,.550$], SE = .024, $p < .001$; $\text{Mean}_{\text{Random List–MDS}} = .501$, [$\text{CI}_{95\%} = .426,.577$], SE = .027, $p < .001$; $\text{Mean}_{\text{Ordered list–Random 2D}} = .454$, [$\text{CI}_{95\%} = .373,.534$], SE = .029, $p < .001$; $\text{Mean}_{\text{Ordered list–MDS}} = .472$, [$\text{CI}_{95\%} = .398,.545$], SE = .027, $p < .001$). As is also visible in this graph, there was a unique trend in the Random List display whereby correct responses were associated with a higher proportion of NN moves ($\rho = .20$ [$\text{CI}_{95\%}$:.095,.301], $p < .001$, $N = 336$). This is consistent with the notion that the correct responses under the Random List

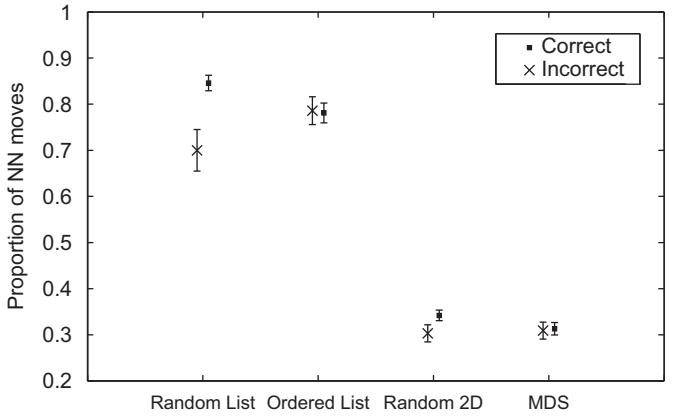

Fig. 14. *Experiment II*: proportion of nearest neighbor (NN) moves split by accuracy of response across the four conditions. One standard error is shown about the mean.

display were associated with participants clicking on adjacent documents. In other words, the increased accuracy under this condition may have been linked to the frequent reliance on the default search strategy (i.e., navigating the visualization via NN moves) when interacting with this particular display.

3.3. Summary

As with the first experiment, the structured 2D display (MDS) performed better than the Random List condition with participants accessing 7 fewer documents and taking 30 less seconds per question. Proportionally, this amounted to 26% fewer documents and 21% less time. However, in contrast to the first experiment, there were differences in accuracy between the different displays. Taking into account both speed and accuracy, MDS was still superior. Although participants were most accurate using the Random List display, this was achieved at the expense of long response times. Analysis of the jumps participants made between document representations in the displays suggested that, on correct trials, participants more often relied on moves between adjacent documents. Such a strategy is ideal in this display. Since the layout is random, the probability that the desired document is directly adjacent to the current document is the same as for any other position on the list. In addition, by clicking on the adjacent representation, the participant is minimizing the effort required to select the next document because the distance needed to move the mouse is kept to a minimum.

In terms of overall performance, it was difficult to distinguish between the two random displays. On the one hand, participants were less accurate on the Random 2D display (by 10%) and less confident (by just under half a point on a seven point scale). On the other hand, they were faster (by 21 s) and accessed fewer documents per question (3.5 fewer) than under the Random list display. Therefore, in cases where there is no inherent structure in the display, showing document representations in two dimensions rather than one did not improve performance.

However, when the displays were structured, the 2D visualization (MDS) outperformed the 1D greedy nearest neighbor algorithm in terms of documents accessed (4.4 fewer per question), with non-statistically significant advantages in terms of accuracy (4%) and speed (19.15 s). This makes sense from a theoretical point of view because the semantic structure can be better represented in two dimensions than one. In summary, when there is content to depict, the 2D representation outperformed the list. However, there was no such difference in the displays of different dimensionality when there was no structure portrayed. In other words, the 2D layout by itself does not assist users in navigating the document space. Rather it is the combination of 2D layout and the faithfulness of this layout in representing a 'human' document space that made the difference.

4. Comparison between experiments

There was a consistent trend across the three visualizations assessed in the two experiments in this paper and in the experiment in Butavicius and Lee (2007). MDS outperformed the Ordered List, while the Ordered List was superior to the Random List. However, the performance advantage was expressed differently between the studies in terms of either speed or accuracy. Fig. 15 shows the relative performance of the three common visualizations in terms of speed and accuracy.

Fig. 16 demonstrates that some of the variability in performance between corpora can be described by a speed-accuracy tradeoff. That is, the different performance between corpora may relate to a change in emphasizing either speed or accuracy at the expense of the other. The correlation between the response time and accuracy across the experiments was .708 ($N=12$, $CI_{95\%}$: .226, .912).

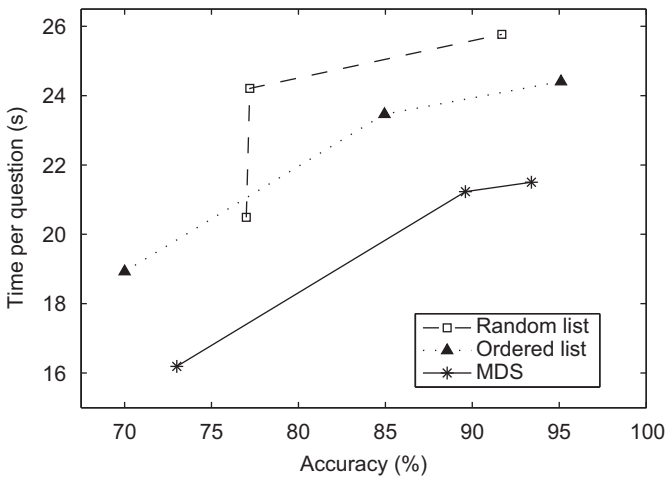

Fig. 15. Comparison of speed and accuracy for the three common visualizations across the three experiments. The bottom right corner of the figure represents ideal performance where participants are both fast and accurate.

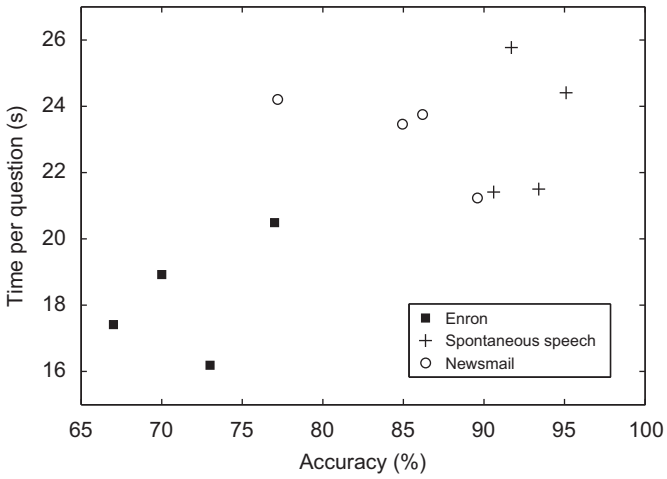

Fig. 16. Performance means for each visualization condition in all three experiments. The symbol indicates which corpus was visualized. The bottom right corner of the figure represents ideal performance where participants are both fast and accurate.

While participants were most accurate when analyzing the spontaneous speech corpus, they also took the longest time to answer the questions. Conversely, those who analyzed the Enron corpus were fastest but this came at the expense of the poorest accuracy across all corpora. The newsmail corpus occupies a position between these two extremes with both accuracy and response times falling between those of the spontaneous speech and Enron sets.

The ease with which information can be found within documents may have produced these effects. For example, it may be that participants found the Enron corpus particularly difficult to understand, and this lowered their expectations on finding relevant information, resulting in the premature ending of searches and guessed decisions. However, many variables differed between these experiments, so a more definitive answer regarding this speed-accuracy tradeoff requires a separate experiment involving different corpora with a (preferably) within-subjects design.

5. Conclusion

In two studies, we found that the 2D visualizations structured according to a cognitive representation of the underlying document similarities outperformed a 1D visualization of the same similarities when applied to unstructured texts. Both of these types of displays performed better than an unstructured list. These findings parallel those for visualizations of highly structured news articles (Butavicius and Lee, 2007). In the second experiment of this paper we also showed that the cognitive representation of the document space was a necessary part of the 2D visualization. Without this structure performance fell to a level similar to a random 1D list.

Across the experiments in this paper and the study in Butavicius and Lee (2007), we found that, in general, the relative performance differences between the visualizations were stable across corpora of different styles. This included well edited news articles, email texts and spontaneous conversational transcripts. Some of the variation in performance between the different corpora may be explained by a change in the tradeoff between speed and accuracy in accessing information.

In addition, this tradeoff may vary across visualizations. In the second study, there was evidence that, in the face of an unstructured list of documents, users could respond more accurately than with the other displays. However, this was at the cost of speed with participants taking the longest to find answers under this condition. Interestingly, the correct responses under this display were associated with navigating the array by clicking neighboring document representations. This type of navigation approach is a brute force method for finding documents in an unstructured list. It may reduce individual mouse movements and guarantee that the user eventually finds the required document(s) but it does not compensate for the display's lack of structure. However, the very fact that participants were still able to find the required documents accurately in the unstructured list should not be discounted and suggests that such displays could be of benefit in tasks where accuracy is more important than speed. Therefore, in the context of the ongoing debate in the literature over whether list or spatialized displays are superior (e.g., Butavicius and Lee, 2007; Cribben and Chen, 2001; Hornbæk and Frøkjær, 1999; Swan and Allan, 1998), we suggest that neither display type is universally advantageous but that different applications favor different tools. For example, an analyst may still find it advantageous to opt for a list-based display in a task where the importance is on accuracy in finding the required information rather than the time it takes to find it.

It should be noted that the generalizability of our findings may be linked to the complexity of the document sets that were visualized. In our study, the document sets appeared to lend themselves to being represented in a 2D space. However, the same may not be true of more complex document and data sets that we may wish to visualize in the real-world. In other words, the success of the visualization approaches tested depends on whether the underlying semantic space of the documents can be meaningfully represented in only two dimensions.

While this study has demonstrated an advantage in cognitively-structured proximity-based visualizations, further research is needed to examine their utility in other real-world applications. For example, a 2D visualization of a complex document space may be particularly beneficial in identifying overall trends in the space. In this case, accuracy is less important because the search is not for a specific document but broader document classifications. Alternatively, when searching for a document, particularly when specific words or terms are likely to be present, visualization may be inferior to keyword or entity-based searches. In addition, as discussed previously, there are a number of other tools that support alternative

investigation of these corpora based on time-line, patterns of correspondence, sentiment and other metadata. Much consideration in any operational scenario has to go into the specific problems that will benefit from a visualization approach and how these approaches can be integrated with more traditional search techniques.

Finally, additional research is required into the specific visual aspects of data visualization. Notably, experiments by Brusco (2007), Fabrikant et al. (2004) and Montello et al. (2003) have begun to directly investigate and model the structure users perceive in point arrays. This research has demonstrated that the structures that users see may be not just those associated with pairwise distances between the points but that other perceptual phenomenon such as emergent features and the vertical illusion may influence their perception of the displays (Montello et al., 2003). Therefore, more fundamental research is necessary to model how users perceive the displays and how our perception deviates from the underlying document space approximated by the visualization.

Acknowledgments

We wish to thank Chlöe Mount, Joanne Spadavecchia and Andrew Brolese for conducting the experiments, Chris Jones for his work on the visualization interface and Ian Coat, Glen Smith and several anonymous reviewers for their assistance and helpful suggestions. Daniel Navarro was supported by an Australian Research Fellowship (ARC Grant DP-0773794).

References

Ashby, F.G., Maddox, W.T., Lee, W.W., 1994. On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. Psychological Science 5 (3), 144–151.

Basalaj, W., 2000. Proximity Visualization of Abstract Data. Unpublished Doctoral Dissertation. University of Cambridge Computer Laboratory.

Brusco, M.J., 2007. Measuring human performance on clustering problems: some potential objective criteria and experimental research opportunities. Journal of Problem Solving 1 (2), 33–52.

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.

Clavier, S.M., El Ghaoui, L.M., 2008. Breaking world news: the computerized dynamic visualization of aggregate perceptions, public opinion, and the making of foreign policy. In: Proceedings of the ISA's 49th Annual Convention, Bridging Multiple Divides. Hilton, CA.

Cohen, J., 1988. Statistical Power Analysis for the Behavioural Sciences second edition Lawrence Erlbaum Associates, Hillsdale, N.J.

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

Cribbin, T., Chen, C., 2001. Visual-spatial exploration of thematic spaces: a comparative study of three visualization models. Electronic Imaging 2001: Visual Data Exploration and Analysis VIII, 199–209.

Cselle, G., Albrecht, K., Wattenhofer, R., 2007. Buzztrack: topic detection and tracking in email. In IUI'07: Proceedings of the 12th International Conference on Intelligent User Interfaces. Honolulu, Hawaii, USA, pp. 190–197.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., Plaisant, C., 2007. Discovering interesting usage patterns in text collections: integrating text mining with visualization. In: Laender, A.H.F., Falcão, A.O. (Eds.), Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. ACM, Lisboa, Portugal, pp. 213–221.

Donath, J., Karahalios, K., Viegas, F., 1999. Visualizing conversation. In: Nunamaker, J.F., Jr. (Ed.), Proceedings of the 32nd Hawaii International Conference on System Sciences. IEEE Computer Society, Maui, Hawaii, pp. 1–9.

Fabrikant, S.I., 2002. Spatial Metaphors for Browsing Large Data Archives. Ph.D. Thesis. University of Colorado-Boulder, Boulder, CO.

Fabrikant, S.I., Montello, D.R., Mark, D.M., 2006. The distance-similarity metaphor in region-display spatializations. IEEE Computer Graphics and Applications 26 (4), 34–44.

Fabrikant, S.I., Montello, D.R., Ruocco, M., Middleton, R.S., 2004. The distance-similarity metaphor in network-display spatializations. Cartography and Geographic Information Science 31 (4), 237–252.

Fortuna, B., Mladenic, D., Grobelnik., M., 2006. Visualization of text document corpus. Informatica 29, 497–502.

Frau, S., Roberts, J.C., Boukhelifa, N., 2005. Dynamic coordinated email visualization. In: Vacla Skala (Ed.), Proceedings of WSCG05—13th International Conference on Computer Graphics, Visualization and Computer Vision. Plzen, Czech Republic, pp. 187–193.

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.

Gersh, J., Lewis, B., Montemayor, J., Piatko, C., Turner, R., 2006. Supporting insight-based information exploration in intelligence analysis. Communications of the ACM 49 (4), 63–68.

Godfrey, J.J., Holliman, E., 1997. SWITCHBOARD-1 Transcripts LDC93S7-T. CD-ROM. Linguistic Data Consortium, Philadelphia.

Görg, C., Stasko, J., 2008. Jigsaw: investigative analysis on text document collections through visualization. In: Attfield, S., Baron, J.R., Mason, S., Oard, D.W. (Eds.), Proceedings of DESI II: Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery. UCL Interaction Centre, University College, London, pp. 59–68.

Granitzer, M., Kienreich, W., Sabol, V., Andrews, K., Klieber, W., 2004. Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In: Proceedings of the IEEE Symposium on information Visualization, pp, 127–134.

Gregory, M.L., Payne, D., McColgin, D., Cramer, N., Love, D., 2007. Visual analysis of weblog content. In: Proceedings of International Conference on Weblogs and Social Media '07. Boulder, Colorado, U.S.A.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101 (Suppl. 1), 5228–5235.

Heeman, P.A., Allen, J.F., 1997. Intonational boundaries, speech repairs, and discourse markers: modelling spoken dialog. In: Proceedings of the 35th Annual meeting of the Association for Computational Linguistics, Madrid, Spain. Association for Computational Linguistics, Morristown, NJ, USA, pp. 254–261.

Hornbæk, K., Frøkjær, E., 1999. Do thematic maps improve information retrieval? In: Sasse, A., Johnson, C. (Eds.), Proceedings of the Seventh IFIP Conference on Human–Computer Interaction (INTERACT'99). IOS Press, pp. 179–186.

Hornbæk, K., Frøkjær, E., 2003. Reading patterns and usability in visualizations of electronic documents. ACM Transactions on Computer–Human Interaction 10 (2), 119–149.

Klimt, B., Yang, Y., 2004. The Enron corpus: a new dataset for email classification research. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (Eds.), Proceedings of the European Conference on Machine Learning, Pisa, Italy. Springer, Berlin, pp. 217–226.

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

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.

Lee, M.D., Pope, K.J., 2003. Avoiding the dangers of averaging across subjects when using multidimensional scaling. Journal of Mathematical Psychology 47, 32–46.

Levelt, W.J.M., 1983. Monitoring and self-repair in speech. Cognition 14, 41–104.

Liu, Y.-H., Dantzig, P., Sachs, M., Corey, J.T., Hinnebusch, M.T., Damashek, M., Cohen, J., 2000. Visualizing document classification: a search aid for the digital library. Journal of the American Society for Information Science 51 (3), 216–227.

Montello, D.R., Fabrikant, S.I., Ruocco, M., Middleton. R.S., 2003. Testing the first law of cognitive geography on point-display spatializations. In: Kuhn, W., Worboys, M., Timpf, S., (Eds.), Proceedings of the Conference on Spatial Information Theory: Foundations of Geographic Information Science (COSIT '03). Springer, Verlag, pp. 316–331.

Morse, E., Lewis, M., 1997. Why information visualizations sometimes fail. In: Proceedings of the 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.

Mothe, J., Chrismenta, C., Dkakia, T., Dousseta, B., Karouacha, S., 2006. Combining mining and visualization tools to discover the geographic structure of a domain. Computers, Environment and Urban Systems 30 (4), 460–484.

Newby, G.B., 2002. Empirical study of a 3D visualization for information retrieval tasks. Journal of Intelligent Information Systems 18 (1), 31–53.

Pirolli, P., Card, S., 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: Proceedings of 2005 International Conference on Intelligence Analysis. MITRE, McLean, VA.

Perer, A., Shneiderman, B., Oard, D.W., 2006. Using rhythms of relationships to understand email archives. Journal of the American Society of Information Science and Technology 57 (14), 1936–1948.

Perer, A., Smith, M.A., 2006. Contrasting portraits of email practices: visual approaches to reflection and analysis. In: Celentano, A., Mussio, P. (Eds.), Proceedings of the Working Conference on Advanced Visual Interfaces AVI '06, New York, USA. ACM Press, New York, pp. 389–395.

Pérez-Quiñones, M.A., Kavanaughm, A., Murthy, U., Isenhour, P., Godara, J., Lee, S., Fabian, A., 2007. VizBlog: a discovery tool for the blogosphere. In: Cushing, J.B., Pardo, T.A. (Eds.), Proceedings of the Eighth Annual International Conference on Digital Government Research: Bridging Disciplines and Domains, Philadelphia, Pennslyvania, USA. Digital Government Society, pp. 314–315.

Powell, L.A., 2004. Visualizing co-occurrence structures in political language: content analysis, multidimensional scaling, and unrooted cluster trees. Journal of Political Language 1 (4) 〈http://www.jdlonline.org/I4Powell1.html〉 (retrieved 29.04.09).

Ratté, S., Njomgue, W., Ménard, P.-A., 2007. Highlighting document's structure. World Academy of Science, Engineering and Technology 31, 34–38.

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., Moorhead, R.J., 2009. A user study to compare four uncertainty visualization methods for 1D and 2D datasets. IEEE Transactions on Visualization and Computer Graphics 15 (6), 1209–1218.

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., 1980. Multidimensional scaling, tree-fitting, and clustering. Science 210, 390–398.

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

Shiffrin, D., 1987. Discourse Markers. Cambridge University Press, New York.

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

Stasko, J., Görg, C., Liu, Z., Singhal, K., 2008. Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization 7, 118–132.

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.

Toffler, A., 1970. Future Shock second ed. Pan Books Ltd., London.

Tory, M., Möller, T., 2004. Human factors in visualization research. IEEE Transactions on Visualization and Computer Graphics 10 (1), 1–13.

Tory, M., Sprague, D.W., Wu, F., So, W.Y., Munzner, T., 2007. Spatialization design: comparing points and landscapes. IEEE Transactions on Visualization and Computer Graphics 13 (6), 1262–1269.

Vickers, D., Butavicius, M.A., Lee, M.D., Medvedev, A., 2001. Human performance on visually presented travelling salesman problems. Psychological Research (Psychologische Forschung) 65, 34–45.

Voorhees, E.M., Harman, D., 2005. TREC: Experiment and Evaluation in Information Retrieval. Cambridge. MIT Press, Massachusetts.

Walenstein, A., 2002. Cognitive Support in Software Engineering Tools: A Distributed Cognition Framework. Unpublished Doctoral Dissertation. Computing Science Department, Simon Fraser University.

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

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.

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.

White, R.W., Muresan, G., Marchionini, G., 2006. Evaluating exploratory search systems. In: White, R.W., Muresan, G., Marchionini, G. (Eds.), Proceedings of the ACM SIGIR '06 Workshop on Evaluating Exploratory Search Systems, Seattle, Washington, USA. ACM, New York, pp. 1–2.

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.