An assessment of email and spontaneous dialogue visualizations

Marcus A. Butavicius [a,b*], Michael D. Lee [c], Brandon M. Pincombe [a]

Louise G. Mullen [a], Daniel J. Navarro [b], Kathryn M. Parsons [a], Agata McCormac [a]

[a] *Defence Science and Technology Organisation, Edinburgh, Australia, 203L, PO Box 1500,Edinburgh, SA, 5111, Australia*

{marcus.butavicius; brandon.pincombe; louise.mullen; kathryn.parsons; agata.mccormac}@dsto.defence.gov.au

[b] *School of Psychology, University of Adelaide, Australia, Adelaide, SA 5005, Australia*

daniel.navarro@adelaide.edu.au

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

mdlee@uci.edu

Running title: Assessing visualizations of unstructured text

[*] Corresponding author. Tel.: +61-8-8259-6097; fax: +61-8-8259-6328.
Postal address: 203L, DSTO, PO Box 1500, Edinburgh, SA, 5111, Australia.

Abstract

Two experiments were conducted examining the effectiveness of visualizations of unstructured texts that consisted of transcriptions of unrehearsed dialogue and emails respectively. In general, the findings of both studies were similar to those of Butavicius and Lee's (2007) experiment which used highly structured news articles; namely, an advantage in semantically structured two-dimensional (2D) spatial layouts such as MDS (multidimensional scaling) over structured and non-structured list displays. The second study 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 (2007) 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 dialogue 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 navigate 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 dialogue

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. Such tools are particularly beneficial when the number of documents in the set to be analysed 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, defence 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. In one such approach, known as a spatial visualization, each document is represented 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. The success of these displays possibly lies in the ability of the human visual system to easily detect 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 optimisation

problems for which the human visual system appears to be very well adapted (see also Vickers et al., 2001).

The reliance of spatial 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. There have been some empirical studies taking up this challenge (e.g., Butavicius and Lee, 2007; Tory and Möller, 2004; Lee et al., 2003; Ware, 2000; Westerman et al., 2005; Westerman and Cribbin, 2000), but, overall, only 3% of the articles on information visualization surveyed in Tory and Möller's (2004) literature review included user studies.

In one user study relating to the current work, Butavicius and Lee (2007) considered the performance of 80 participants in an experiment using four different visualization techniques. 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 this 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) spatial displays outperformed the one-dimensional (1D) lists.

As with most user studies in visualizations, Butavicius and Lee (2007) used documents that were well-edited in the form of news articles. 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 language texts. Spontaneous speech, by its very nature, is less structured and offers different challenges to analysis (for both humans and machines) than written or prepared speech. Linguistic features such as speech repairs (Levelt, 1983) and discourse markers (Shiffrin, 1987) can make interpretation of such language difficult (Heeman and Allen, 1997), as can the unique and often fluid vocabulary of such texts.

In addition to traditional spontaneous speech, there has been a rapid increase in the use of internet forums, web logs, chat rooms, email communications and instant messaging (Lyman and Varian, 2003). In such fora, the language is also more spontaneous, conversational and less polished and the vocabulary and linguistic style vary from those found in more formal texts. For example, such texts often use slang terms, unique abbreviations and constrictions and may contain many misspellings. The texts also range widely in the degree of formality and length. Finally, many such texts are not self-contained and form part of an ongoing discussion (Perer and Shneiderman, 2005).

The very large document sets provided by these newer forms of spontaneous speech can be valuable sources of information. Extracting information from these corpora is the sort of problem for which visualization and other analytical tools could be of great benefit. An ethnographic study by MacKay (1998) revealed that different people use

varying methods for managing their own email archive. In addition, email is used not just as a means of communication (both formal and informal), but also as a tool to manage time, information and tasks. As a result, it is unlikely that any one singular approach to the visual analysis of email corpora is optimal. Rather, different approaches suit different analyses and this variety is reflected in the tools for email analysis described in the literature. Perer, Shneiderman and Oard (2006) have demonstrated the insights made possible by analysing the temporal rhythms of correspondence in emails using only header information. In this approach, context to conversations is provided by analysing the pattern of activity for a relationship within an individual's or community's email archive. Similarly, Perer and Smith (2006) received favorable subjective assessments from eight participants when they were presented with Correspondent Treemaps (displaying hierarchical information within an email archive), Correspondent Crowds (displaying correlational information demonstrating mutuality and balance between a user and their correspondents) and Author Lines (displaying temporal rhythms of initiation and reply). Other email visualization interfaces such as MAILVIEW (Frau et al., 2005) allow users to navigate email repositories within time and date plots using coordinated views while Görg and Stasko (2008) have demonstrated the application of the JIGSAW tool to visualise an email corpus with an emphasis on representing the entities represented in the texts and their relationships.

There has also been research in the area of visualising other less formal, unstructured texts in the form of web logs (colloquially known as "blogs"). Using the INSPIRE tool, Gregory et al. (2007) constructed spatial visualizations of blogs (referred to as a "Galaxy view") to explore semantic content. The underlying semantic similarity

judgements were provided by Latent Semantic Analysis (Deerwester et al., 1990), a technique which is able to model human judgments of similarity very well for some texts (Lee et al., 2005). The IN-SPIRE tool also supports complex querying and affect analysis in an attempt to examine the sentiments that frame blog statements. Similarly, Pérez-Quiñones et al. (2007) have developed the VizBlog tool to visualize the structure of conversations and content similarities across blog entries. Other tools, such as CONVERSATIONAL LANDSCAPE and LOOM have focussed on representing the communication patterns of text-based conversations from newsgroups and web forums (Donath et al., 1999).

Despite this increased interest, we are not aware of a user study that has sought to evaluate how well visualization tools assist a user in the analysis of the new forms of language found in these communication domains. In this article, 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 snapshot of texts across a number of individuals. This represents the scenario where an analyst, for the purposes of business, political or security intelligence gathering, is surveying current opinions or facts from a corpus of unstructured, spontaneous texts from multiple authors. As a result, there was never more than one extract from a thread or ongoing discussions in any of the test sets we used. 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).

We were implicitly testing how faithful the visual representation of the document space was to the user's expectations of semantic similarity. This was achieved by using actual 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) (for further discussion see Butavicius and Lee, 2007). By focussing on the 'visual' rather than the 'cognitive' component of a visualization, our approach is similar to the *defeatured* systems approach of Morse and Lewis (Morse and Lewis, 1997; Morse et al., 2002) with the exception that we are testing with an empirical psychology paradigm.

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). However, while the previous study used a between-subjects design, the current experiment used a repeated measures approach. Although this required multiple document sets, it provided greater statistical power for the same number of participants.

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.

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 dialogues were not scripted but were spontaneous.

For the purposes of the current project, all documents were processed to remove notation indicating non-linguistic utterances and sound. Document excerpts were selected to conform to a hierarchical taxonomy. Specifically, the topics were arranged into five categories (SPORTS, CRIME, CARS, POLITICS and MISCELLANEOUS), each of which contained a number of topics, as shown in Figure 1. In the first four categories the documents were all semantically related to documents within the same category. In the OTHER category the documents were such that none of them were judged by the authors to be 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.

< INSERT FIGURE 1 ABOUT HERE >

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), and did not require a high-level interpretation of the

document. In many real-world scenarios, there are a number of other techniques besides document visualization, such as keyword search, that could be more effective in finding such information. However, the questions in this study are an experimental tool used to indicate how well the user can identify trends in the document space (e.g., to find clusters of associated documents as well as outlier documents) in a manner that does not involve extensive interpretation of the documents. Such additional interpretation would involve higher cognitive and linguistic abilities, which would interact with the primary focus of the experiment, namely, how well the displays present information visually. It should be noted, however, that interpretation of the concepts and topics contained in the documents is still a large part of the user's interaction with the visualization display in this experiment. Theoretically, the greater the consistency between the semantic content and the visual organisation of documents in the interface, the more effective the visualization.

Different types of questions were included to allow examination of whether the visualization techniques benefited specific types of information retrieval 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 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

An example of a question from the fourth document set in which participants were required to find two documents from the same topic (BASEBALL) in the same category (SPORTS) is:

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 is the second option (B). The order of the response options was randomized for each trial.

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 that provided all pairwise judgments and used the set of judgments from the one participant who demonstrated the greatest variation in assigning similarity scores.

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.

[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 optimisation of these displays for visualization see Butavicius and Lee (2007).

< INSERT TABLE 1 ABOUT HERE >

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

< INSERT FIGURE 2 ABOUT HERE >

< INSERT FIGURE 3 ABOUT HERE >

2.1.5. Interface

The interface is shown in Figure 4. The top right pane contains the visualization of the corpus. The color of the icons indicates the status of the document representation with

respect to the search actions that have been completed for that particular question. Specifically, colors indicate whether the document representation had been accessed (blue) or not (tan) and also which document representation is currently active (green). After the participant has answered the question and provided a confidence rating the color of all the document icons reset to tan for the following question. The background of the visualization was white.

< INSERT FIGURE 4 ABOUT HERE >

The text of the active document is displayed in the top left pane. Directly below the visualization and the document pane is the question pane (tan). Underneath this are 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 require participants to provide a response option before the confidence ratings, the confidence ratings were inactivated until a response option was selected. At this point the response options and document icons were inactivated until a confidence response had been selected.

2.1.6. Experimental design

All participants were presented with each of the four different display types where each of these displays visualized one of the four different document sets. 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 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.

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. 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 Figure 5, there is a trend visible indicating a speed advantage for the 2D structured visualizations, especially over the Random List condition. Bonferonni multiple comparisons indicated a significant difference between the Random List and ISOMAP displays ($Mean_{difference} = 25.61$s, $CI_{95\%}$= [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 ($Mean_{difference} = 26.16$s, $CI_{95\%}$= [-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 seconds 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.

< INSERT FIGURE 5 ABOUT HERE >

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, $F(3,45) = 11.461$, $p < .0001$) with a large effect size (multivariate η_p^2 =.433). Bonferonni comparisons confirmed the trend visible in Figure 6 – that fewer documents were accessed using the structured visualizations compared to Random List ($Mean_{\text{Random List – Ordered List}} = 4.799$, $CI_{95\%}$= [.771, 8.826] SE = 1.462, $p = .012$; $Mean_{\text{Random List – MDS}} = 7.625$, $CI_{95\%}$= [3.844, 11.406] SE = 1.372, $p < .001$; $Mean_{\text{Random List – ISOMAP}} = 7.431$, $CI_{95\%}$= [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.

< INSERT FIGURE 6 ABOUT HERE >

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 an ideal 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 minimises 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.

Sequentially accessed documents can either be nearest neighbors or not and the proportions of these are displayed in Figure 7 for each visualisation. As expected from Figure 7, the proportion of nearest neighbor (NN) moves varied significantly between the displays (Wilks' Lambda = .168, $F(3,45) = 74.13$, $p < .0001$) 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 Bonferonni comparisons (given in Table 2) were significant and those between 1D and 2D displays were significant at $\alpha = 0.001$.

< INSERT FIGURE 7 ABOUT HERE >

< INSERT TABLE 2 ABOUT HERE >

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 Bonferonni 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 Figure 8 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 η_p^2 = .102).

< INSERT FIGURE 8 ABOUT HERE >

While some question types were more or less difficult than others, the actual display condition did not improve or decrease performance between different questions. Rather, it influenced performance over all the questions to a similar effect. 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.

< INSERT TABLE 3 ABOUT HERE >

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 is a large positive correlation between the time taken to respond to a question and the number of documents accessed (ρ= .799 [$C1_{95\%}$: .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 (ρ= -.145 [$C1_{95\%}$: -.201,-.088], p < .001, N = 1152) and reduced confidence (ρ= -.141 [$C1_{95\%}$: -.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 (ρ=.461 [$C1_{95\%}$:.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 [$C1_{95\%}$:.039, .265], p < .001, N = 1152; $\rho_{\text{ordered list}}$

=.14 [$C1_{95\%}$:.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.

2.3. Summary

In summary, users performed better with the structured 2D visualizations than the random list approach. They were 25 seconds faster and accessed 5-8 fewer documents per question. Proportionally, this amounted to 17% less time and 17 - 27% fewer documents. 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 produced distinctly different interpretations of the semantic structure of the corpora, there was no significant performance difference between them.

3. Experiment II: Enron Emails

Experiment II differs from Experiment I in two main ways. Firstly, the document set consisted of emails from the Enron Corporation rather than transcriptions of spoken dialogue. 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 has not been 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 document 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. 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 it's dealings 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 FERC (Federal Energy Regulatory Commission) 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 organised for Enron employees.

For the NON-WORK area the topics included:

- **Charity Events** (2) – Charity events organised 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 less 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

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.

3.1.5. Interface

The interface was identical to that used in Experiment I.

3.1.6. Experimental design

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 Figure 9. The RMANOVA indicated that there was significant variation in accuracy associated with the different displays (Wilks' Lambda = .828, $F(3, 46) = 3.176$, $p < 0.05$, 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 2D random display compared to the random list ($Mean_{\text{Random 2D – Random List}} = 9.9\%$, $CI_{95\%} = [0.7,19.1]$ SE = 0.033, p = .028).

< INSERT FIGURE 9 ABOUT HERE >

However, examination of the response times suggests that, overall, performance on the Random List display was 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, $F(3, 46) = 3.106$, $p <$

0.05, multivariate $\eta_p^2 = .168$) and, as can be seen in Figure 10, 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 ($Mean_{Random\ List - MDS} = 30.11$, $CI_{95\%} = [0.20,60.02]$ SE = 10.87, $p < .05$). This amounts to an average reduction in time taken of 21%. Overall, the mean response time was 127.78 seconds, with a standard deviation of 103.85 seconds.

< INSERT FIGURE 10 ABOUT HERE >

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 1.25% of the variation ($\rho = -.112$ [$C1_{95\%}$: -.164,-.059], $p < .001$, $N = 1344$). There was a significant difference in the number of documents accessed between the displays (Wilks Lambda = .643, $F(3, 46) = 8.522$, $p < 0.001$, multivariate $\eta_p^2 = .357$). As is visible in Figure 11, Bonferonni comparisons revealed that MDS was associated with significantly fewer documents accessed than the random list ($Mean_{MDS - Random\ List} = -7.55$, $CI_{95\%} = [-11.62,-3.47]$ SE = 1.482, $p < .001$). This amounts to a reduction of 26% in the number of documents accessed compared to the random list.

< INSERT FIGURE 11 ABOUT HERE >

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 (ρ = 0.61 [CI $_{95\%}$: 0.58, 0.64], $p < 0.001$, $N = 1344$). 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 Figure 12, the list displays appeared to attract a higher proportion of NN moves than the 2D displays and this was supported by the Bonferonni comparisons ($Mean_{Random\ List - Random\ 2D} = 0.483$, [CI $_{95\%}$= 0.416,0.550] SE = 0.024, p < .001; $Mean_{Random\ List - MDS} = 0.501$, [CI $_{95\%}$= 0.426,0.577] SE = 0.027, p < .001; $Mean_{Ordered\ list - Random\ 2D} = 0.454$, [CI $_{95\%}$= 0.373,0.534] SE = 0.029, p < .001; $Mean_{Ordered\ list - MDS} = 0.472$, [CI $_{95\%}$= 0.398,0.545] SE = 0.027, 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 (ρ= 0.20 [CI $_{95\%}$: 0.095,0.301], $p < 0.001$, $N = 336$). This is consistent with the notion that the correct responses under the random list 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.

< INSERT FIGURE 12 ABOUT HERE >

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 – because 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 seconds) 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 (3.5%) and speed (19.15 seconds). Theoretically, it seems likely that more of the cognitive structure can be represented in two dimensions than one. An additional factor in our study is that, unlike the 2D display, the 1D display represents only ranked distance data. 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 to portray. In other words, the 2D layout by itself does not assist users in navigating the document space. 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. Figure 13 shows the relative performance of the three common visualizations in terms of speed and accuracy.

< INSERT FIGURE 13 ABOUT HERE >

Figure 14 demonstrates that some of the variability in performance between corpora can be described by a speed-accuracy tradeoff (i.e., that the different performance between corpora may relate to a change in bias towards either speed or accuracy). The correlation between speed and accuracy across the experiments was .708 (N = 12, $CI_{95\%}$: 0.226, 0.912). While participants were most accurate when analysing the spontaneous speech corpus, they also took the longest time to answer the questions. Conversely, those who analysed 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.

< INSERT FIGURE 14 ABOUT HERE >

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 data, resulting in participants prematurely ending searches and making guesses. 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 this study, 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 change in a bias towards either speed or accuracy in accessing information.

In addition, this bias may also vary across visualizations. In the second study, there was evidence that, in the face of an unstructured list of documents, users could still respond accurately. 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 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.

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 particular 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.

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 J. 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 & Hall, London.

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

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

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.

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. Philadelphia: Linguistic Data Consortium.

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.

Gregory, M.L., Payne, D., McColgin, D., Cramer, N., Love, D., 2007. Visual analysis of weblog content. In: 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.

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.

Lyman, P., Varian, H.R., 2003. How much information 2003. Retrieved Jul 19, 2005 from <http://www.sims.berkeley.edu/how-much-info-2003>.

MacKay, W., 1998. More than just a communication system: Diversity in the use of electronic mail. In: Greif, I. (Ed.), Proceedings of the 1998 ACM Conference on Computer-Supported Cooperative Work, ACM, NY, pp. 344-353.

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.

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.

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., 2005. Beyond threads: identifying discussions in email archives. Technical Report, Maryland University College Park, Human Computer Interaction Lab, ADA440462.

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 8th Annual International Conference on Digital Government Research: Bridging Disciplines & 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)). Retrieved on April 29 2009 from <http://www.jdlonline.org/I4Powell1.html>.

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.

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

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

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.

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.

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

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	*	*

Table 2.

Bonferonni comparisons for the proportion of nearest neighbor moves

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

* and ** indicate significant differences at the .05 and .001 alpha levels respectively

Table 3.

RMANOVA

Source	Measure	df (error)*	F	p	η_p^2
Question	Accuracy	5(235)	3.316	0.011	0.066
	Docs accessed	5(235)	14.022	<0.001	0.230
	Confidence	5(235)	4.998	0.001	0.096
	Time	5(235)	15.324	0.000	0.246
	Prop NNs	5(235)	1.192	0.116	0.039
Question x Display	Accuracy	15(705)	1.472	0.165	0.030
	Docs accessed	15(705)	1.164	0.314	0.024
	Confidence	15(705)	0.892	0.533	0.019
	Time	15(705)	1.125	0.343	0.023
	Prop NNs	15(705)	.988	0.451	0.021

*All df's used in calculations corrected using the Greenhouse-Geisser technique.

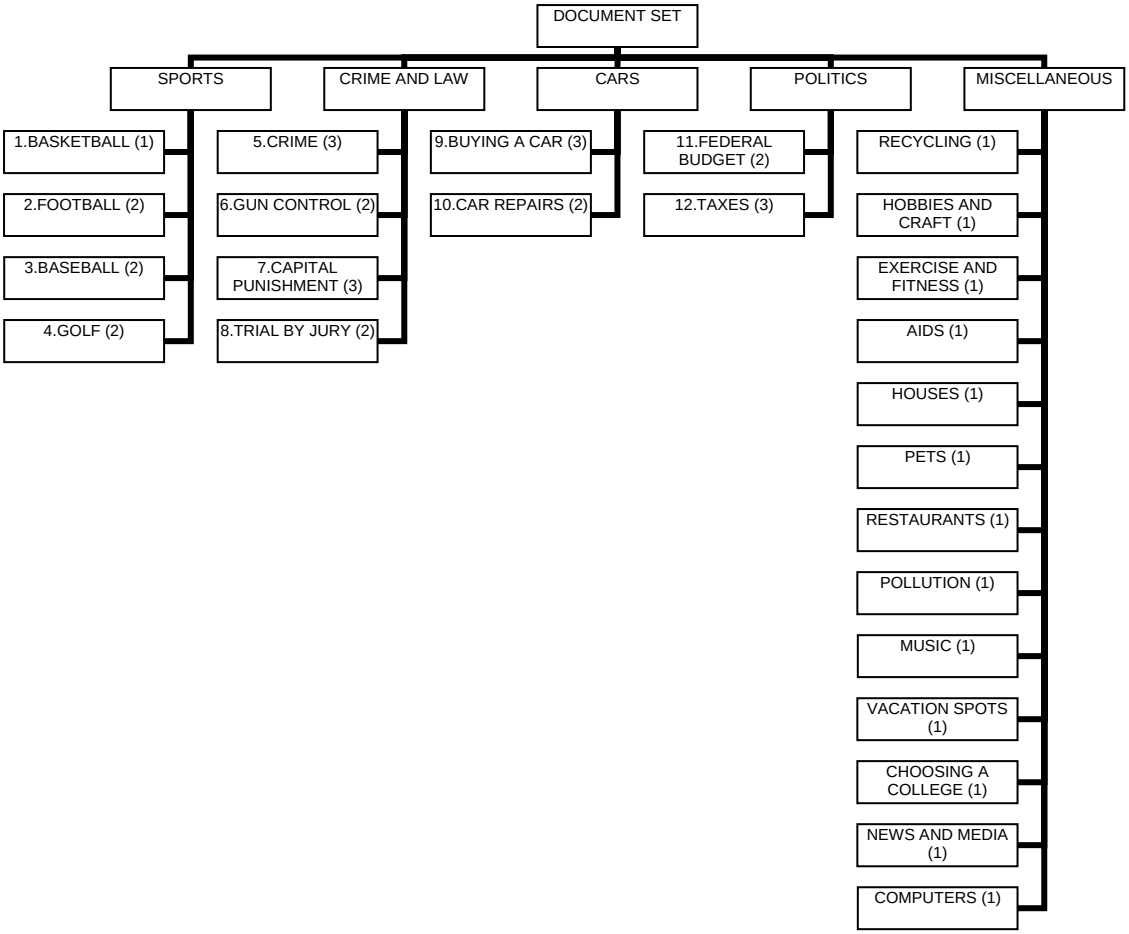

Figure 1. The arrangement of topics into categories. The number preceding all topic names from the semantically coherent categories indicate the graph labels for Figures 2 and 3. The number of documents for each topic is shown in brackets. This structure was identical for all four test document sets.

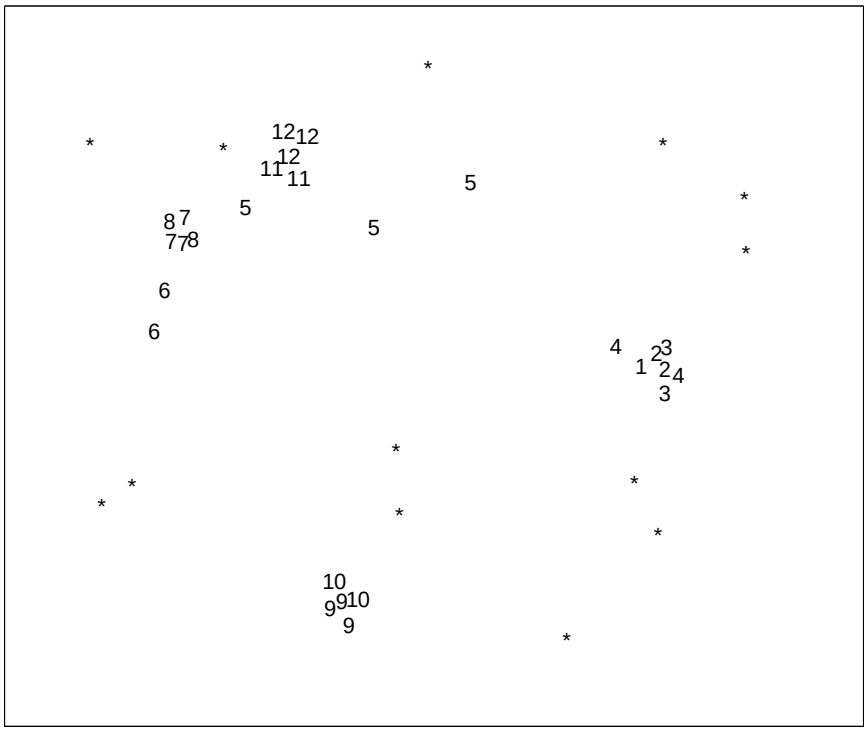

Figure 2. Representation of the MDS solution for the first document set. The topic membership is indicated by a number for documents belonging to semantically coherent categories and by an asterisk for those in the ‘Miscellaneous’ category. The graph labels are contained in Figure 1.

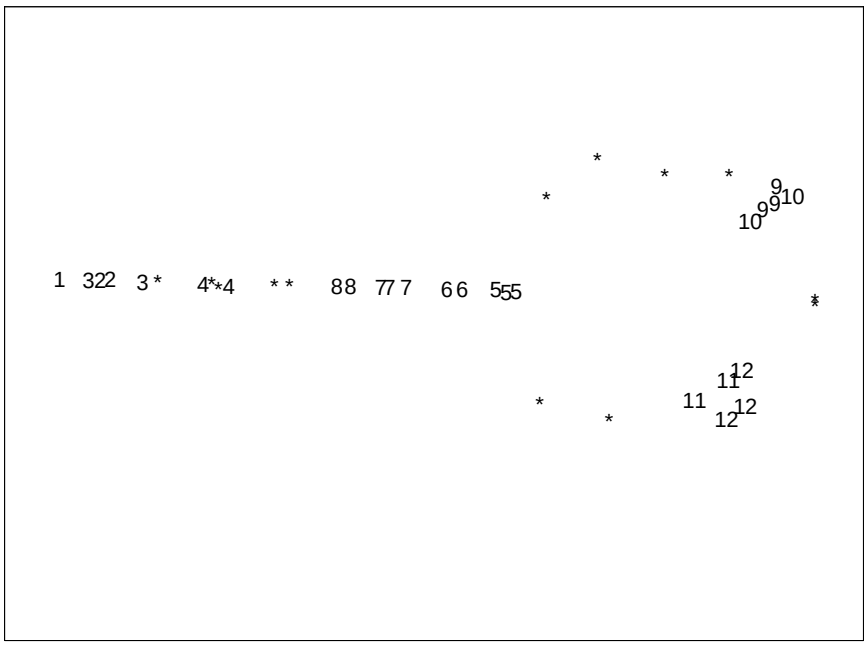

Figure 3. Representation of the Isomap solution for the first document set. The topic membership is indicated by a number for documents belonging to semantically coherent categories and by an asterisk for those in the ‘Miscellaneous’ category. The graph labels are contained in Figure 1.

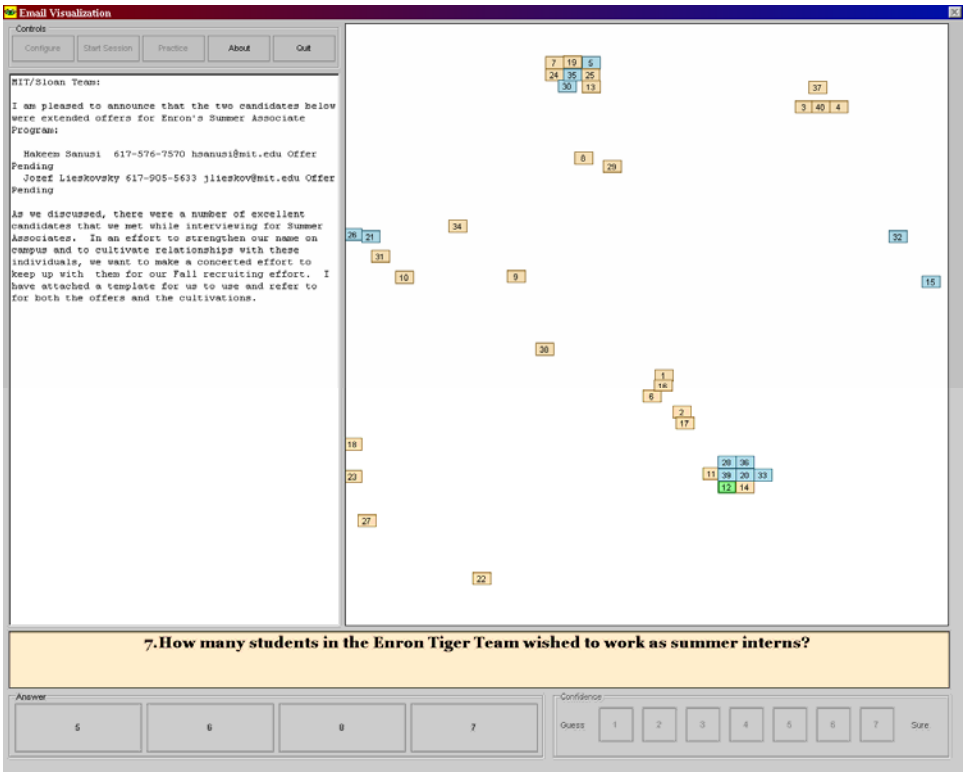

Figure 4. Screenshot of the interface showing an MDS 2D visualization of one of the Enron email document sets.

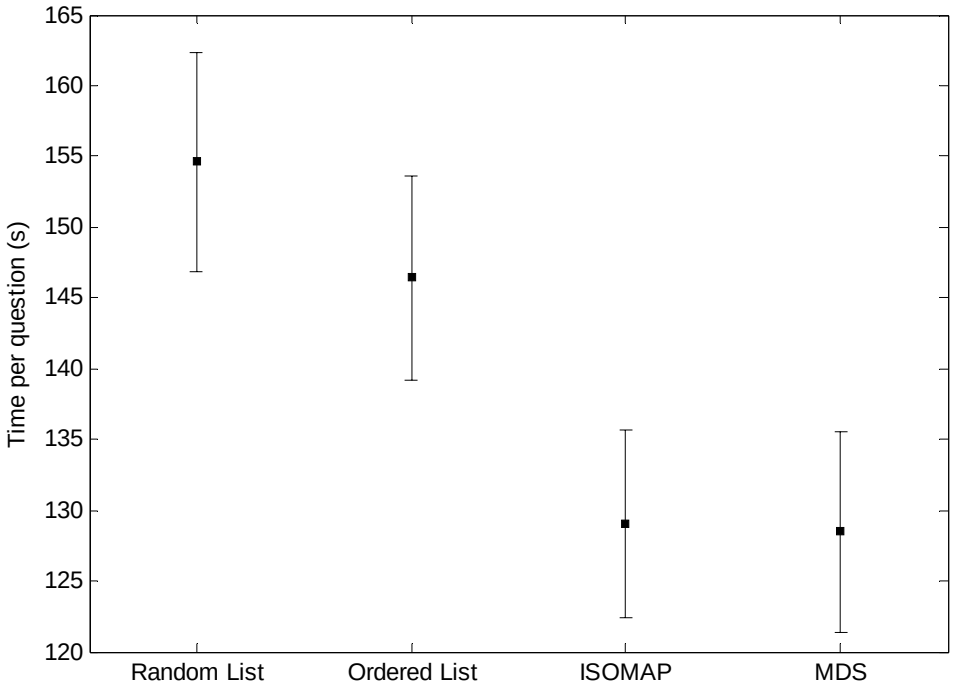

Figure 5. Mean response time (s) across the four conditions. One standard error is shown about the mean.

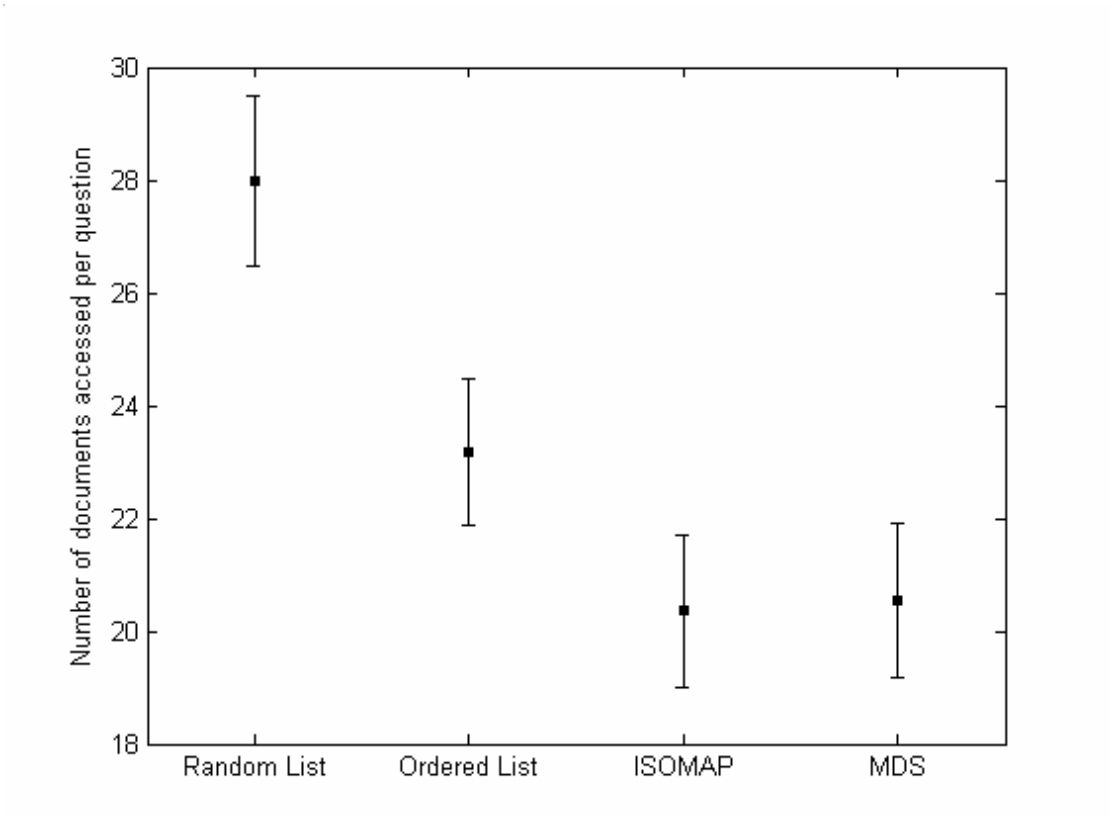

Figure 6. Mean number of documents accessed per question across the four conditions. One standard error is shown about the mean.

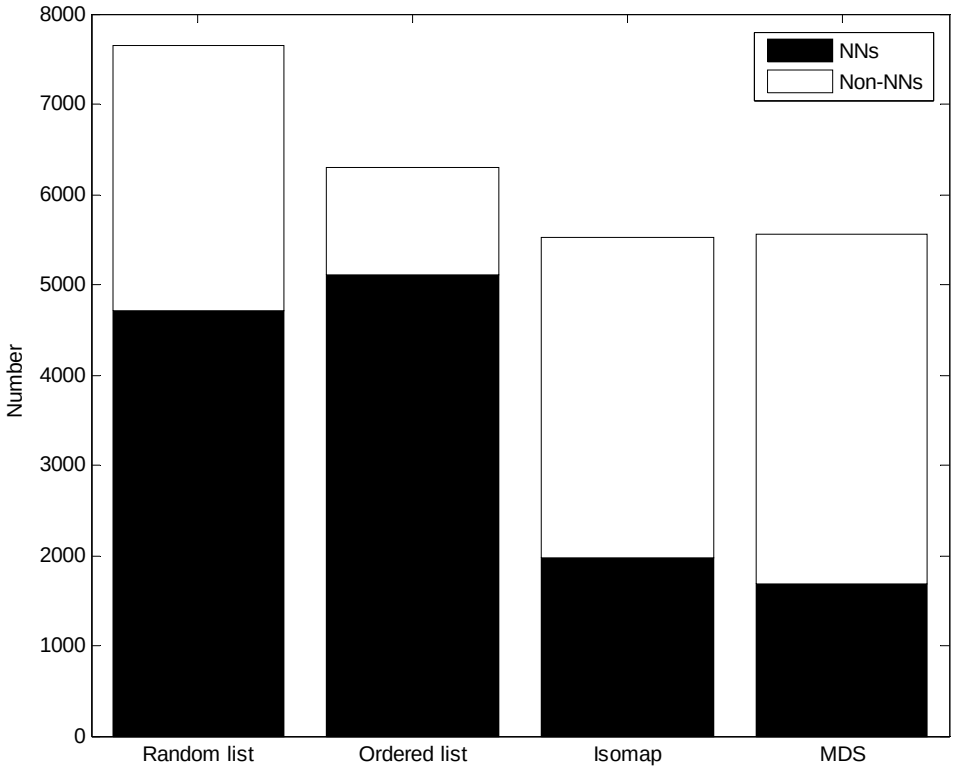

Figure 7. Stacked bar graphs of the number of moves made by participants in the display across the four conditions and classified by nearest neighbor (black) and non-nearest neighbor moves (white).

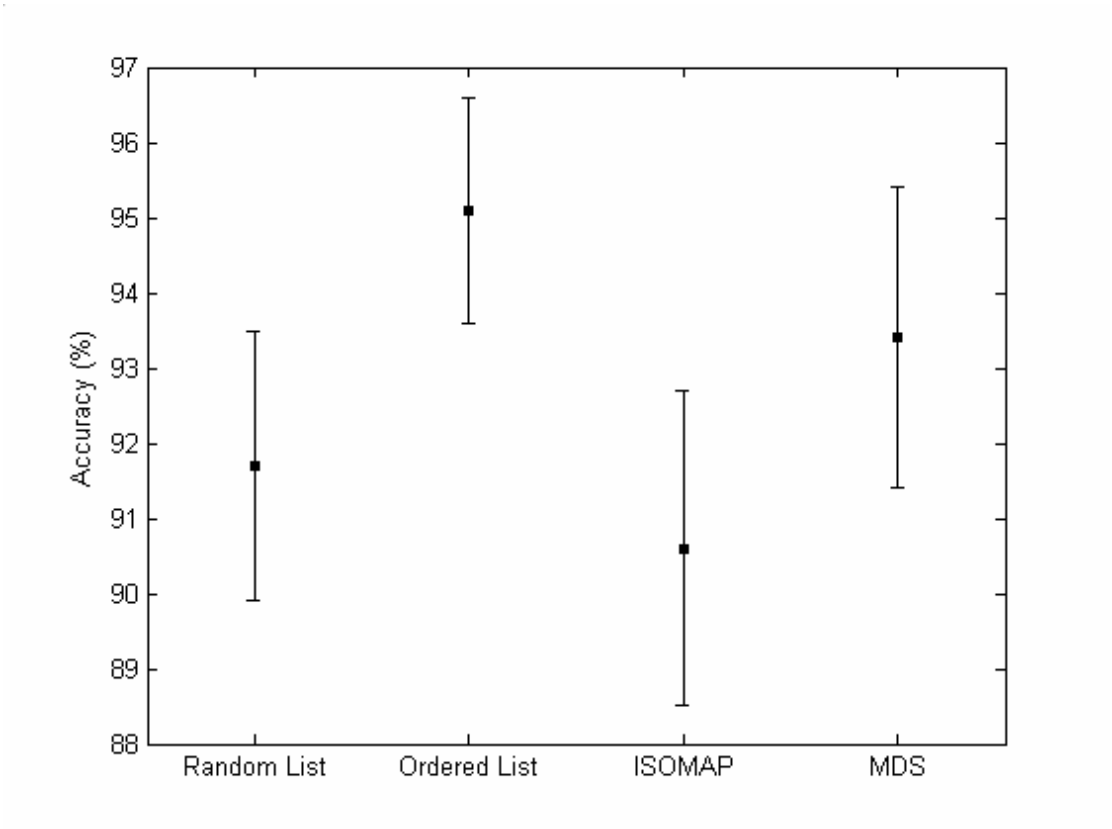

Figure 8. Mean accuracy scores across the four conditions. One standard error is shown about the mean.

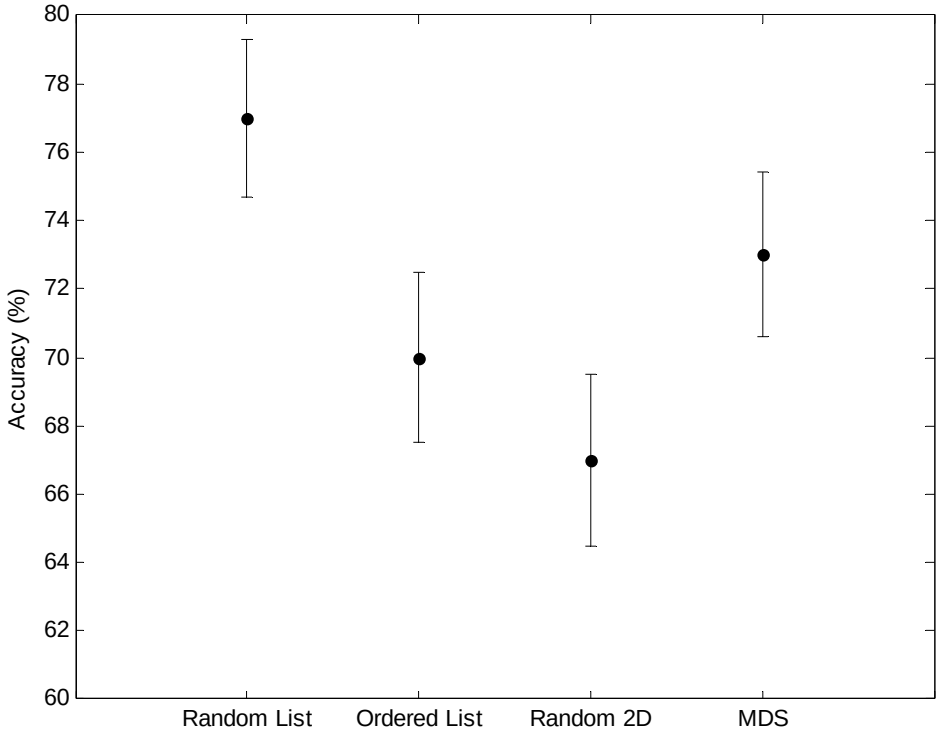

Figure 9. Mean accuracy across the four conditions. One standard error is shown about the mean.

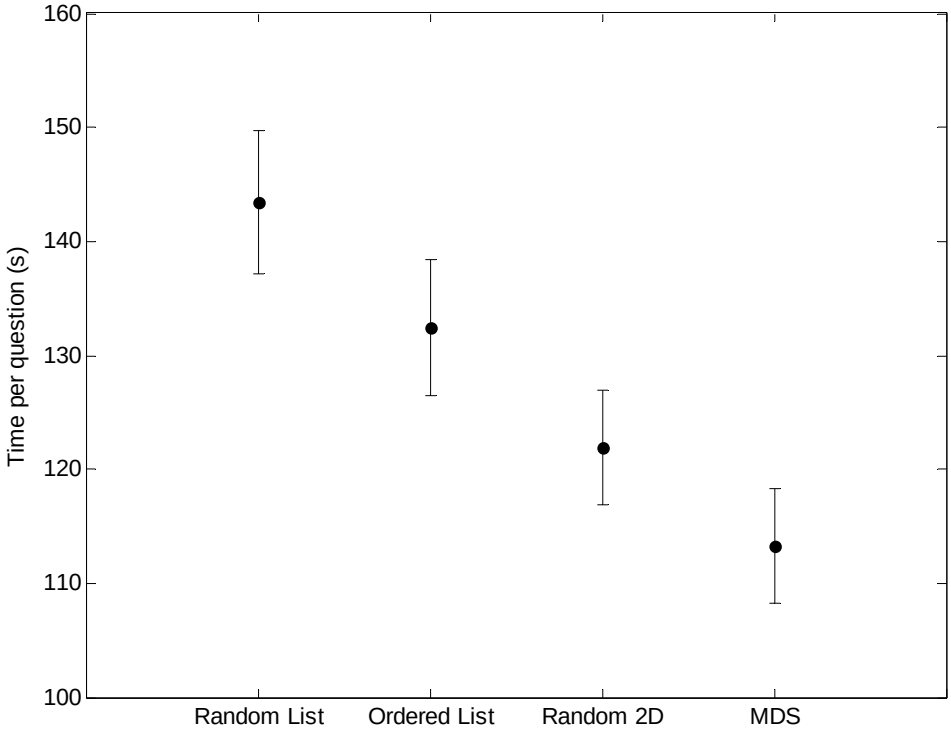

Figure 10. Mean response time (s) across the four conditions. One standard error is shown about the mean.

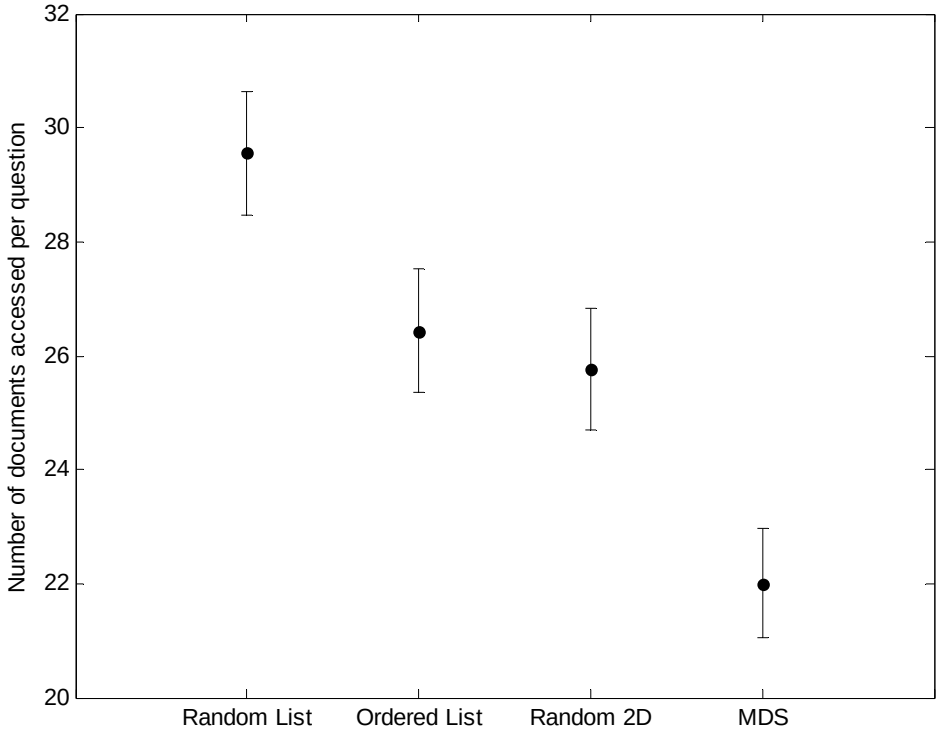

Figure 11. Mean number of documents accessed per question across the four conditions. One standard error is shown about the mean.

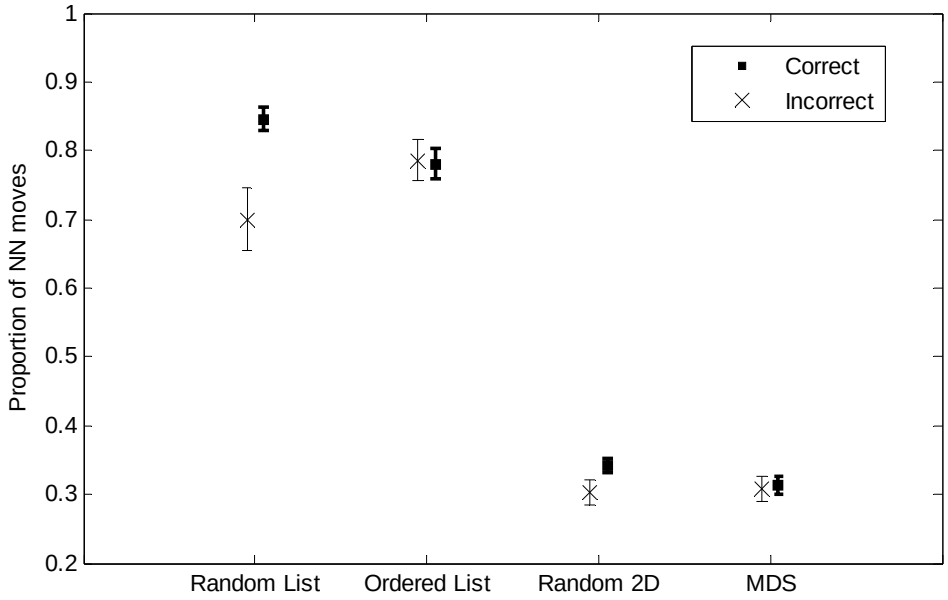

Figure 12. Proportion of nearest neighbour (NN) moves split by accuracy of response across the four conditions. One standard error is shown about the mean.

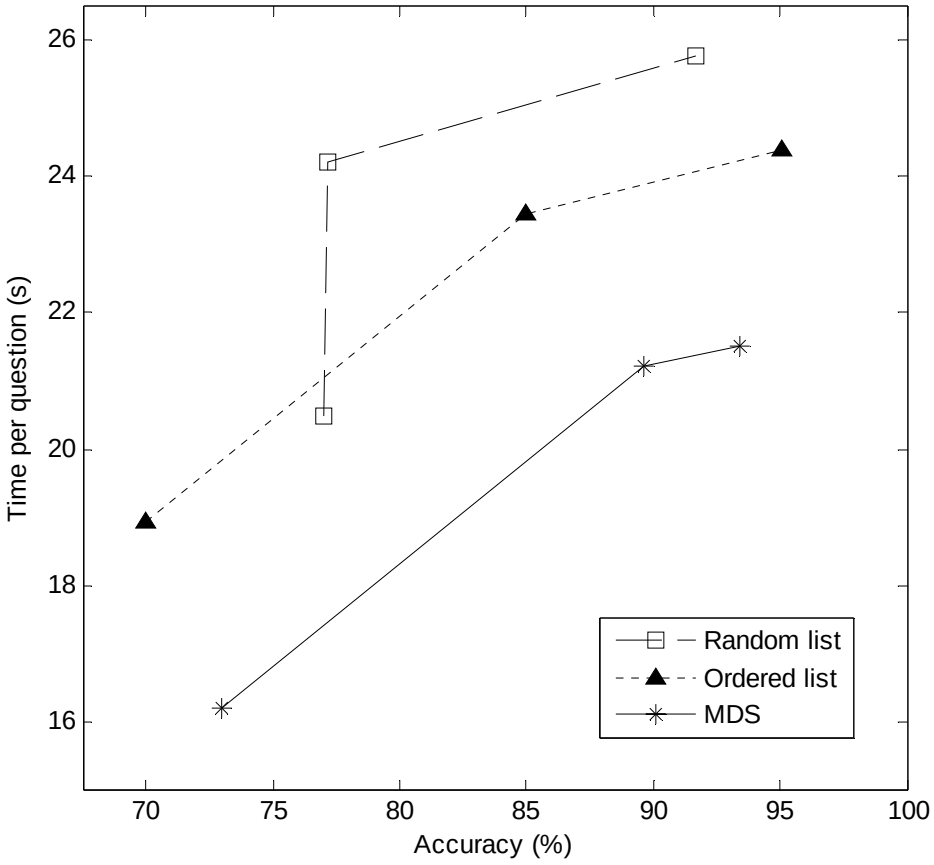

Figure 13. Comparison of speed and accuracy for the three common visualizations across the three experiments. The bottom right corner of the figure represents ideal performance where participants are both fast and accurate.

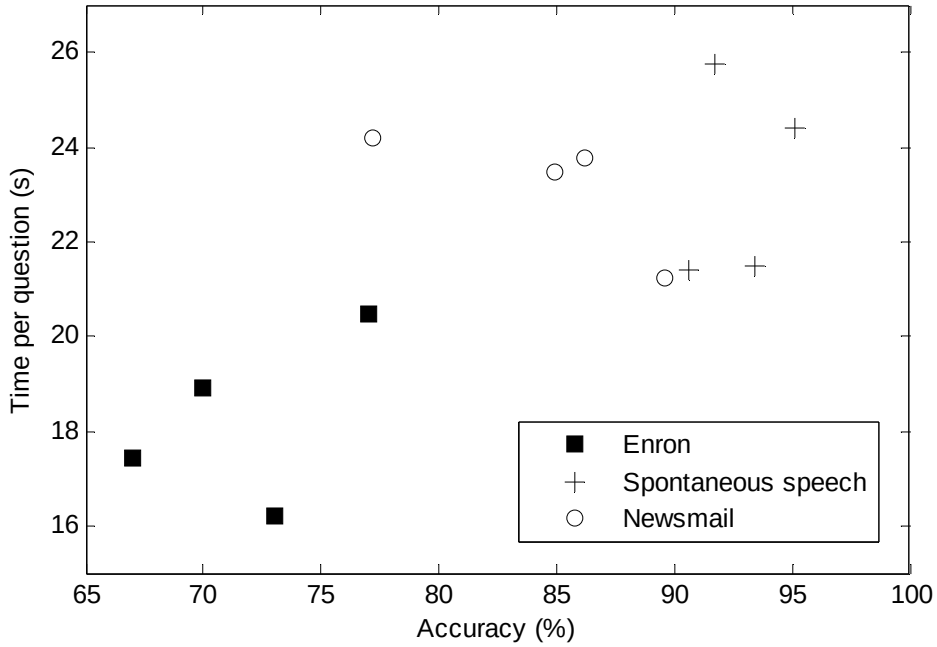

Figure 14. Performance means for each visualization condition in all three experiments. The symbol indicates which corpus was visualized. The bottom right corner of the figure represents ideal performance where participants are both fast and accurate.