Int. J. Human-Computer Studies (2000) **53,** 741–763
doi:10.1006/ijhc.2000.0416
Available online at http://www.idealibrary.com on IDEAL®

Evaluating the effectiveness of visual user interfaces for information retrieval

A. G. Sutcliffe
Centre for HCI Design, Department of Computation, UMIST, P.O. Box 88, Manchester M60 1QD, UK

M. Ennis
Centre for HCI Design, School of Informatics, City University, Northampton Square, London EC1V 0HB, UK

J. Hu
Centre for HCI Design, Department of Computation, UMIST, P.O. Box 88, Manchester M60 1QD, UK

(Received 31 January 2000, and accepted in revised form 31 May 2000)

An integrated visual thesaurus and results browser to support information retrieval was designed using a task model of information searching. The system provided a hierarchical thesaurus with a results cluster display representing similarity between retrieved documents and relevance ranking using a bullseye metaphor. Latent semantic indexing (LSI) was used as the retrieval engine and to calculate the similarity between documents. The design was tested with two information retrieval tasks. User behaviour, performance and attitude were recorded as well as usability problems. The system had few usability problems and users liked the visualizations, but recall performance was poor. The reasons for poor/good performance were investigated by examining user behaviour and search strategies. Better searchers used the visualizations more effectively and spent longer on the task, whereas poorer performances were attributable to poor motivation, difficulty in assessing article relevance and poor use of system visualizations. The bullseye browser display appeared to encourage limited evaluation of article relevance on titles, leading to poor performance. The bullseye display metaphor for article relevance was understood by users; however, they were confused by the concept of similarity searching expressed as visual clusters. The conclusions from the study are that while visual user interfaces for information searching might seem to be usable, they may not actually improve performance. Training and advisor facilities for effective search strategies need to be incorporated to enhance the effectiveness of visual user interfaces for information retrieval.

1. Introduction

Few empirical studies of visual user interfaces (UIs) have focused on the strategies that users employ when searching through or interacting with complex images. We do know that the mapping between visual representations and the user mental model is important for promoting comprehension and effective use of visual UIs, but there is little

1071-5819/00/110741 + 23 $35.00/0

information on how visual structures provide cues for learning. One solution is to provide users with facilities for organizing their own maps of information spaces. Usability studies have shown that self-organization of visual mental maps is effective (Czerwinski, Van Dantzich, Robertson & Hoffman, 1999); however, in many cases the user may have no starting point for self-organization, so the system needs to present a learnable image that portrays the structure of an underlying system or database. While many designs have been developed to provide visualizations of hierarchies, networks and time lines, evaluation studies are less common. Some metaphors have demonstrated effectiveness, such as the time line metaphor (Plaisant, Milash, Rose, Widoff & Shneiderman, 1996), and simple categorial grouping of results in information retrieval searches (Pirolli, Schank, Hearst & Diehl, 1996). While a development framework for design of visualizations to match the users' task and underlying system structures has been proposed (Card, Mackinlay & Shneiderman, 1999), questions still remain about the effectiveness of visualizations in specific formats or metaphors in supporting particular tasks.

In our previous work, we investigated user interaction with 3D information browsing displays (Sutcliffe & Patel, 1996) and demonstrated that 3D images do not appear to provide a significant advantage over two-dimensional displays. We also investigated the strategies users employed while interacting with 3D information visualizations and showed that different patterns of visual search range from systematic searches following the image structure to random sampling. In subsequent studies on standard GUI-based information-retrieval systems we found that users' search strategies were a major determinant of search performance and this seemed to over-ride usability problems (Sutcliffe, Ryan, Springett & Doubleday, 2000). One of the major determinants of search success was the user's persistence in iterative cycles of search and careful evaluation of retrieved articles to determine their relevance. The other major factor was choice of appropriate keywords. Furthermore, search facilities needed to be closely coupled so users could dynamically explore the relationship between query terms and the retrieved documents. The effectiveness of close coupling has been demonstrated in the alphaslider system (Ahlberg & Shneiderman, 1994), which integrated visualization and interactive querying; however, these designs did not support query articulation or exploration of meta-data, and were limited to query templates with value ranges. Other prototypes have provided integrated support for both query formulation and results evaluation, e.g. in the OKAPI system (Hancock-Beaulieu, Fieldhouse & Do, 1995) relevance feedback facilities allow the user to extract terms from retrieved documents for reuse in subsequent queries. Although a wide variety of visual user interface designs have been developed for browsing thesaurus and classification structures, visualization of retrieved results is less common. Furthermore, visualization for query formulation and results evaluation does not seem to have been integrated in one system. This forms the starting point for the design we report in this paper.

As well as building a novel visual interface for information searching we were interested in how visual metaphors represent system models to the user. Most visualizations have represented information structures as hierarchies or networks of various forms (Card, Robertson & Mackinlay, 1991). More adventurous representations such as data walls portray design concepts such as filters and context of focus, while tilebar browsers have depicted properties of retrieved documents. In this study, we attempted to

test how visualization might be able to communicate the underlying search functionality of a system. We chose similarity-based searching with the latent semantic indexing algorithm (Landauer & Dumais, 1997) and cluster analysis of retrieved results. This gives a more sophisticated search process than the traditional search engine with keywords and relevance ranking based on goodness-of-fit metrics. The question was how well could visualizations communicate something of the search engine process as well as the structure of the information base.

The paper is organized into five sections. In the next section, we describe the design of the integrated information browser system. This is followed by the methods used in the empirical study on the design to determine its effectiveness and analyse usability problems. We then report performance results and usability problems that were experienced with the design; and users' behaviour and search strategies. The paper concludes with a discussion of lessons learned from the empirical study.

2. Interface design

The Integrated Thesaurus-Results Browser was designed to support cycles of iterative querying, browsing and evaluation of retrieved results. In previous studies (Sutcliffe & Ennis, 1998), we proposed a cognitive task model that described user behaviour during different phases of information searching tasks and the search support facilities required to support each phase. For instance the model indicated that the early phases of articulating search needs and forming queries should be supported by thesauri and term suggestion facilities, while the evaluation phase needed visualization of retrieved results, with sorting and clustering to show grouping and relationships between documents, as well as relevance ranking of articles. Furthermore, our model indicated a close coupling between visualization in the query formulation and results evaluation. These requirements have been discovered and implemented before in isolation, for instance, the scatter/gather browser (Pirolli *et al*., 1996) provides clustering algorithms for grouping results sets and limited visualization, while the information visualizer (Card *et al*., 1991) concentrates on more sophisticated visualization for the query articulation phase.

The Integrated Thesaurus-Results Browser takes these design concepts one stage further. The screen layout is illustrated in Figure 1.

A tiled window layout was adopted to increase the ease of cross referencing between queries, meta-data representation in the thesaurus, visual summaries of the result sets and viewing documents. A tiled display saves the user work and reduces working memory load by allowing the user to continually scan all the necessary information for the task in hand. Inevitably, the downside of this design choice is a more busy display and limited area for viewing large-scale visualizations. The screen is divided into six areas. The top left-hand panel contains the query formulation dialogue, which allows query terms to be entered directly or to be selected from the thesaurus immediately below it. The thesaurus has a standard hierarchical structure, ranging from general to more specialized terms, although synonyms and related terms were not included in the case study system. The thesaurus hierarchy contained 118 terms arranged in six top level categories which expanded into 2–3 lower levels. Since the whole thesaurus was too large to fit into the allocated space, controls allowed the user to expand sub-trees by single clicking upper-level terms. Double-clicking terms entered the term selected as part of the

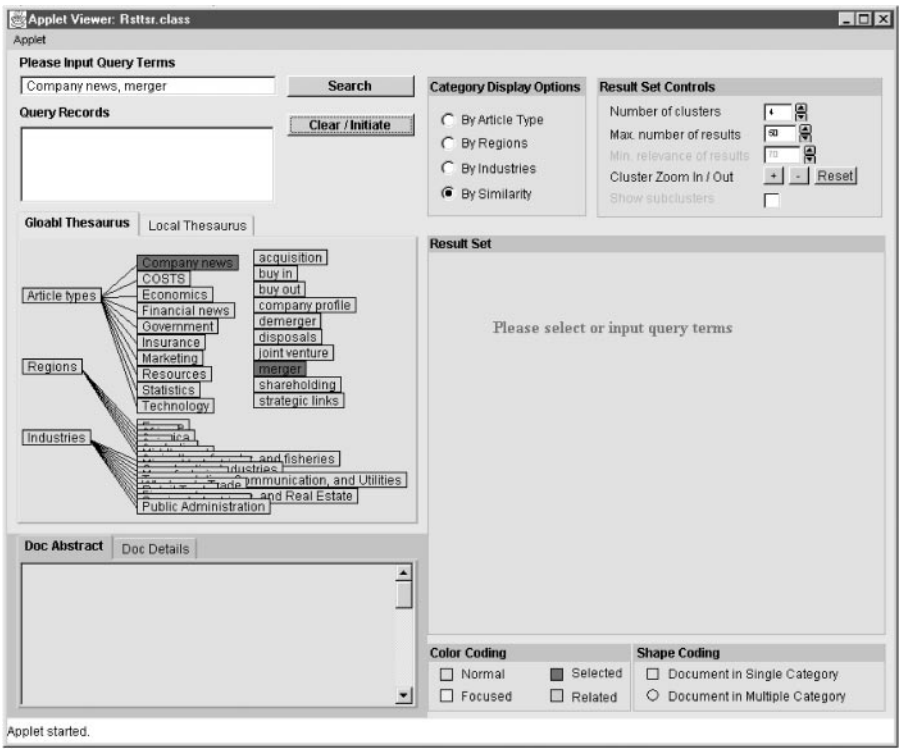

FIGURE 1. Overview of the Integrated Thesaurus-Results Browser system. The user selects terms from the thesaurus (middle left) to form the query (top left). The LSI search algorithm followed by cluster analysis produces the bullseye display (middle right) that gives groups of similar results in each cluster ranked by relevance to the user's query in the target "bullseye" rings. Filters and controls in panel (top right) enable the user to control the number of clusters or selected results grouped by indexing categories. Selecting a document from the bullseye displays the abstract and document details in the results viewer (lower left).

current query. The bottom part of the left side of the screen contained the abstract and details of the retrieved document, which were selected in the results browser on the middle right-hand side of the screen. The results browser contained one or more "bullseye" cluster displays for similar documents with the retrieved results for a query. The bullseye metaphor encoded two properties of the results: first, relevance was represented by rings with higher relevance in the centre of the bullseye, moving outwards to lower scores; secondly, similarity between documents was expressed by clusters on the browser.

The system used the latent semantic indexing (LSI) algorithm (Landauer & Dumais, 1997) as its search engine. LSI works by calculating the similarity between word distributions in two or more documents using an eigenvector algorithm. We used it first to retrieve documents which matched the query keywords using the LSI term to document similarity calculation as a relevance measure. Individual documents within a retrieved set were then matched for similarity using LSI's document to document similarity. In the latter, LSI calculates the co-occurrence between the distribution of all

terms in any two documents, giving a similarity score for each pair of documents. Similarity scores were calculated for all dyadic combinations of documents in a set. A minimum spanning tree cluster analysis algorithm was then applied to the matrix of document to document scores to discover groups of highly related articles. The cluster analysis works by selecting the most closely related document pair, working downwards from those documents to discover nearest neighbour closely related documents and so on. Branching in the tree depends on the number of child documents related to the current "parent". Clusters can be extracted at different levels of similarity score by counting all documents connected to a parent at an arbitrary level.

In the results browser the default cluster display was set at two groups, although this was under user control, and in the case of a tie at the first level (e.g. 3 equal scores for the first set of children documents), the system defaulted to a single group. The number of clusters selected could be set by the user with the control panel on the top right hand of the screen. This also allowed clusters to be selected by a standard indexing method, so documents were grouped by any category used to index the database. The control panel also provided parameters to set the number of results retrieved in relevance order, percentage relevance cut-off, etc.

The results in the bullseye cluster display appear in a spiral from the centre outwards with distance from the centre encoding relevance to the user's query. The bullseyes can be moved and expanded to improve the view when dense clusters of documents are displayed. Moving the cursor over each document symbol causes the short title to be displayed as "hover text" and double clicking on the icon triggers display of the abstract and document details in the results viewer area, as illustrated in Figure 2.

The system was implemented on a sub-set of the Financial Times McCarthy database of newspaper articles. It was programmed in Java and runs under Windows NT on a Pentium P60 using an MS Access database.

3. Experimental methods

Twelve users (5 males, 7 females, age range 23–46) participated in the experiment. The information searching experience of the users, captured by a pre-study questionnaire, is summarized in Table 1. Most users were researchers or students at City University, but they had diverse backgrounds and interests and represented a wide cross-section of casual and professional users.

The study consisted of five phases, organized in the following sequence.

(1) Pre-test questionnaire to capture subject experience and demographics.
(2) System training in which the basic operations of the user interface were explained by running through a typical query and evaluation of results. The concept of similarity-based searching and the metaphors used in the thesaurus and results presentations (bullseyes) were explained. The subjects were given the opportunity to ask questions and were encouraged to try out the system facilities.
(3) Experimental task; the subjects were asked to carry out two searches as follows.
    (a) Please find articles discussing company mergers using the terms available in the thesaurus. Note that it may be necessary to explore different thesaurus terms to do this effectively.

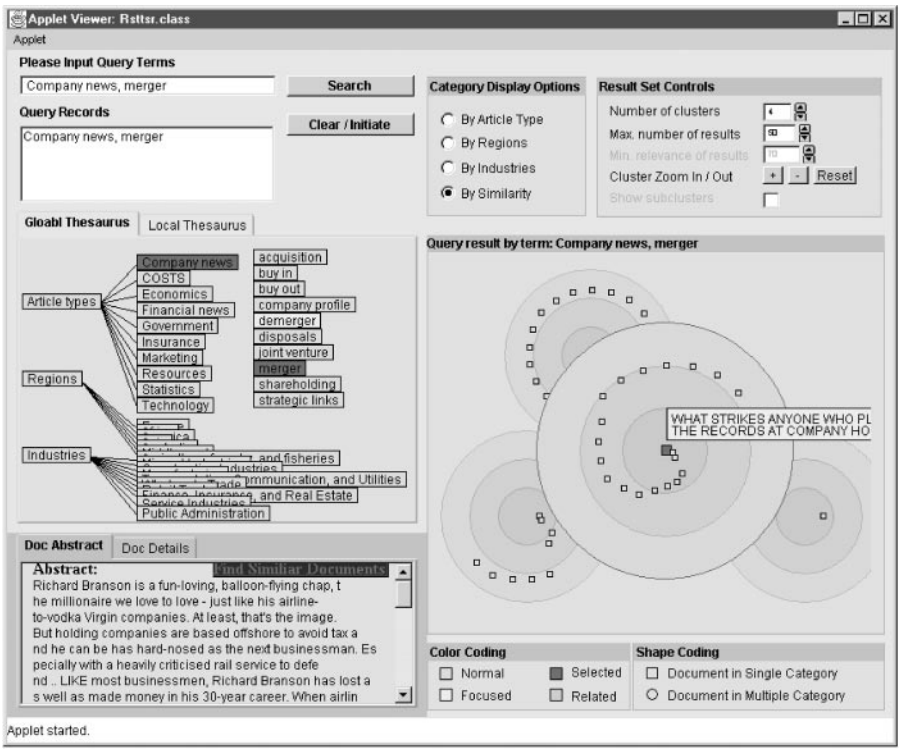

FIGURE 2. The results browser showing "hover text" titles of articles and expanded bullseye display.

TABLE 1
Pre-test questionnaire results on a cued 1–5 *scale* (< 1 *week—many times/day*)

Question	Mean score
1. Frequency of browsing for information in databases	3.1
2. Frequency of searching for information in databases	1.8
3. Frequency of browsing information on the WWW	4.8
4. Frequency of searching for information on WWW	4.0
5. Overall Web usage	4.0

(b) Please find articles which discuss the link between interest rates, inflation and wages.

Both searches were designed to encourage similarity-based searches and required multiple term queries for effective searches. The subjects were encouraged to continue searching as long as they needed to gather all the relevant articles they

thought might be present in the database. The subjects had to browse the results display, identify relevant articles from the hover text, then view the abstract and document details to determine if they were relevant. They gave the article number to the experimenter if it was judged to be relevant. All the retrieved and inspected articles were recorded. During the experiment subjects were encouraged to verbalize any problems they encountered with the system and the experimenter asked follow up questions to clarify the nature of such problems, following the practice of co-operative evaluation (Monk & Wright, 1993). The experimenter helped with usability problems only when users encountered usability breakdowns and could not proceed.

(4) After the experimental task the subjects filled in a post-test questionnaire in which they rated the usability of system facilities on a 1–5 scale.
(5) A de-briefing interview completed the session, in which the experimenter ran through a structured list of questions to probe the subjects' understanding of all the system facilities, metaphors and underlying model (i.e. similarity-based searching). In this phase, the subjects were also asked to interpret a screen diagram of the bullseye metaphor, make suggestions for improvements to the system and explain any particular problems they had experienced.

The subjects were paid £20 for their participation. Sessions lasted between 45 min and $2\frac{1}{4}$ h, with the experimental task durations ranging between 20 min and $1\frac{1}{4}$ h.

The subjects all searched the web frequently; however, database usage (bibliographic and numeric) was less frequent.

Three sets of data were collected. First, performance data for information retrieval were measured as the number of documents indicated by the subjects to be relevant to the query after viewing the abstract or the bullseye display hover text. These documents were compared with an expert's judgement of which documents in the databases were relevant, to determine a % recall measure. The precision of the subjects' retrieval was then calculated from the % of relevant documents within their retrieved set. The second dataset of usability measured the problems encountered by the subjects when using the interface as well as their reported comprehension of user-interface metaphors and functionality. The third dataset recorded subjects' behaviour patterns when interacting with the system, as a set of mental and physical behaviours.

The data were analysed to investigate individual differences and correlations between performance, usability and behaviour, to answer the following questions:

(1) Did poor usability and comprehension of the visual user interfaces result in poor performance?
(2) What patterns of usage behaviour were shown and did these correlate with performance and usability data?

4. Results

4.1. RETRIEVAL PERFORMANCE

Performance was assessed against a gold standard of relevant documents, as judged by an independent expert who was familiar with the domain and read all the articles in the test database. The maximum number of relevant documents for each task was 21 with no

TABLE 2
Recall and precision performance for both tasks by subject

	Relevant documents retrieved		Total retrieved		Recall		Precision	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
					%	%	%	%
	a	*b*	*c*	*d*	$a/21*100$	$b/21*100$	$a/c*100$	$b/d*100$
JU	3	4	15	23	14.3	19.0	20.0	17.4
AG1	3	4	16	18	14.3	19.0	18.8	22.2
HS	3	3	18	18	14.3	14.3	16.7	16.7
MK	2	4	13	21	9.5	19.0	15.4	19.0
KK	2	4	11	19	9.5	19.0	18.2	21.1
AG2	2	3	11	14	9.5	14.3	18.2	21.4
RM	1	3	7	13	4.8	14.3	14.3	23.1
RV	4	0	13	1	19.0	0.0	30.8	0.0
MD	2	1	10	6	9.5	4.8	20.0	16.7
AM	2	1	9	5	9.5	4.8	22.2	20.0
CN	1	1	3	4	4.8	4.8	33.3	25.0
GS	0	1	4	5	0.0	4.8	0.0	20.0
Total	25	29	130	147				
Mean	2.1	2.4	10.8	12.3	9.9	11.5	19.0	18.6

Number of relevant documents for each task = 21.

overlap between tasks, and the database contained a total of 123 documents. Recall is the number of relevant records retrieved as a percentage of the total number of relevant records in the database; precision measures the percentage of retrieved records that are relevant. The performance results are illustrated in Table 2.

Overall performance was poor. The subjects' retrieval of relevant documents ranged from 0 to 4 out of a total of 21 relevant articles in the database. Three subjects managed an average recall score $> 10\%$ for both tasks (subjects JU, AG1, HS). Overall, performance was similar on both tasks although subject RV had reasonable recall and precision in the first task but failed to find any relevant documents in the second. Most subjects selected several documents as being relevant in both tasks, with the exception of subjects CN and RV in task 2. The five worst performers (RV, MD, AM, CN and GS) also selected fewer documents, averaging 7.8 and 4.2 in the respective tasks, compared to the whole group averages of 10.8 and 12.3. Motivation may have had an effect but the subjects were well rewarded for the experiment and did take their duties seriously. Another explanation may lie in lack of domain knowledge for assessing article relevance. The database of Financial Times articles was not the subject area of any of the participants; however, the articles were not technical and an accurate assessment of relevance could have been expected by most educated members of the public.

We had no immediate explanation for the poor recall, although some reasons did emerge later when interaction with the system and usability problems were analysed.

TABLE 3
Post-test questionnaire ratings of the interface design, mean scores on 5-point semantic differential scale (e.g. very confusing … very clear)

Question	Mean score
Presentation and structure of the thesaurus was clear	3.5
Thesaurus helped development of queries	3.4
Easy to navigate around the thesaurus	3.5
Query history was clear	2.6
System easy to use for querying	3.8
Training required for effective use	3.6
Clear relationship between query and bullseye display (results browser)	3.2
Understood the reasons for location of documents in bullseye display	4.0
Visual representation of results was clear	3.6
Bullseye display useful for navigating through results	4.1
Visual representation of results adds value over a list	4.2

4.2. SUBJECTS' RATINGS

In the post-test questionnaires the subjects rated the system favourably in spite of their poor performance. The query facilities and thesaurus (see Table 3) were rated above the mid-range on the 5-point scale, and the overall opinion on ease of querying gained a 3.8 point score. The results browsing and presentation facilities were also favourably rated by the subjects, with the bullseye display in particular receiving a 4-point plus rating. The weaker features were the history of queries (2.6) that could not be inspected easily and the link between queries and the search results display (3.2).

4.3. USABILITY EVALUATION

Usability was assessed by observing difficulties and problems encountered by users which either interrupted normal interaction or caused the user to abandon the current task. There were no occasions when users were so confused that they had to abandon the task, so the following data refer to "critical incidents" in the Monk and Wright (1993) terminology. Errors were categorized following the model mismatch evaluation method (Sutcliffe *et al.*, 2000) which classifies errors to help diagnose the underlying design feature that caused the difficulty.

The top 10 errors with their categories are given in Table 4.

TABLE 4

Usability errors observed during the tasks, categorized by design feature and problem category. Where one user experienced the same error more than once the absolute frequency is given (in parentheses)

	Error-design feature	Number of users	Error category
1	Selecting thesaurus terms	8 (12)	Execution mismatch
2	No history list for previously selected articles	7	Missing functionality
3 =	Confusing representation of the thesaurus hierarchy	6	Misleading cue + metaphor mismatch
3 =	No search progress timer	6	Missing feedback
5	Unpredictable movements in bullseye displays	5 (6)	Execution mismatch
6 =	No labels on bullseye clusters	4	Missing feedback/ functionality
6 =	Unpredictable scroll control in abstracts	4	Inappropriate functionality
8 =	Interaction not possible during search	3	Inadequate feedback
8 =	Truncated text in bullseye hover text	3	Impossible action, missing functionality
8 =	No history; repeat previous searches	3	Missing functionality

The overall frequency of errors was low. Nearly all errors were experienced only once by each user, so overall the system was usable and many of the usability problems observed pointed to missing requirements rather than design defects. The most frequent error, double clicking on the thesaurus terms to select them for a query, was caused by an excessively long inter-click interval so users found rapid double clicks had no effect. The second problem was a missing requirement for an article visit list history because users want to see the documents they had already chosen. The next (3rd =) problem, cluttered thesaurus display, obscured the hierarchical ordering; moreover, the expansion controls made links between hierarchy levels difficult to follow. In the other 3rd = problem the system did not give any feedback when the search was being executed and interaction was not possible during the search (8th = problem). Three users tried to enter another query or interact with the thesaurus map while the search was in progress, but the system only allowed single-threaded queries. The unpredictable movement of the bullseye display (6th =) was a programming error that made the display move in an unpredictable manner. The scroll control in the abstract window was also a simple programming

TABLE 5
Top ten user comprehension and usability problems reported in the post-test interviews

No.	Usability/comprehension problem	No. of users
1	Thesaurus categories and views not clear	9
2	Coding articles shared between clusters not understood	8
3	Thesaurus links between hierarchy levels not clear	7
4	No search progress indicator	7
5	Thesaurus term selection difficult	6
6	No bullseye cluster labels	5
7	Number of clusters not clear	5
8	Abstract and document details not clear	5
9	Whole display too cluttered, hard to comprehend	4
10	No history of previously selected articles	3

problem, as the scroll bar did not reset to the top of an abstract when the user moved from one document to the next. The absence of labels on the bullseye categories (6th =), although infrequent, was symptomatic of deeper problems in understanding the display metaphor. The final two (8th =) errors—truncated text in the bullseye display and no facility to reuse or repeat a previous search—point to a display feedback problem and a missing requirement.

At first sight the usability evaluation gave few reasons why user performance might be poor. The error rate was low, and most problems (see Table 5) were either missing requirements (problems 2, 4, 8 and 10), or simple programming problems (1, 5, 7 and 9), leaving only two problems that were not immediately easy to fix. The thesaurus display problem (5) was partially a consequence of the decision to use a tiled window display for consistency, thus restricting the area of the thesaurus display. However, the display clearly needed considerable improvement with better scaling/zoom controls and clearer representation of the hierarchy. The lack of labels on the bullseye clusters has no immediate solution. The LSI/cluster analysis retrieval process has no means for summarizing or identifying the relatedness of any one cluster. Human intervention is necessary to inspect the cluster of documents and assign a descriptive label summarizing the cluster's raison d'être.

The post-test de-briefing interview concentrated on further diagnosis of observed or reported problems during the session and systematically probed the users' understanding of system functionality, metaphors used in the thesaurus and results browsers and the user's model of similarity-based search. The experimenter followed a structured interview approach and asked probe questions for each area of the system in turn: thesaurus, query

formulation, results browser, abstract viewer and filter controls. Users were also encouraged to make suggestions for improving the system to remedy the problems observed during the experimental sessions. In this phase, more deep-seated problems were discovered that demonstrated that, even though superficially the system appeared to be usable, poor user understanding made usage sub-optimal. The top 10 user-comprehension problems are given in Table 5.

Several problems (1, 3, 4, 5, 6 and 10) reported by the users were identical to those observed during the test sessions, which is not surprising, although the frequency of users reporting the problem was often higher than error frequencies observed during the experimental session. More interesting are the problems not apparent in the evaluation sessions. Articles shared between two or more clusters (2), were represented as a circle whereas mono-cluster articles were shown as squares. This coding was not understood by $\frac{2}{3}$ of the users, possibly because it did not directly impinge on the experimental tasks, even though it was useful for exploring similarity and in assessing document relevance. The number of clusters in the bullseye metaphor was not clear to five subjects. The default was set at two but the users were confused by this and did not understand the system model of similarity. The number of clusters could be restricted to any arbitrary cut off the user wished; in addition, the clusters could be organized into seven industrial sector categories. Most users either stuck with the default two clusters or reduced it to one. The relationship between the abstract and the document details was criticized by five users who preferred to have the document details (author, date, keywords, etc.) first and the abstract second. Finally, four users commented that the whole display was cluttered and remarked that an overlapping rather than a tiled layout might be more effective. Two lower frequency errors not in the top 10 were lack of multi-tasking (reported by two subjects), and absence of highlighting query keywords in abstracts. The frequency of errors by subject showed a fairly even distribution, median 6, range 3–9, so only one subject encountered a high number of problems.

Debriefing interviews uncovered serious problems with the user's comprehension of the system metaphors and search functionality. The misunderstandings of metaphors and system functions that were inferred from subjects' statements in post-test interviews are given in Table 6.

Seven users experienced problems with the thesaurus. Only one user did not understand the basic hierarchy model, but she and six others had problems with the hierarchy levels and links between them, unevenness of the tree (some sub-branches had more sub-levels than others), the choice of categories and the absence of synonyms. Three users wanted to have their own customized thesaurus. Browser controls which determined the number of clusters were poorly understood, and these users also had problems with the enlarge and move functions. The concept of search by example, "find similar to this article", and use of similarity clustering as an aid to results evaluation was used by six subjects, four of whom reported that they did not understand it; a further two did not understand and did not use it. The verbal reports indicated that their model of the system, possibly influenced by experience with Web search engines (see Table 1), was a simple frequency count of keyword hits determining retrieval relevance, rather than the more sophisticated LSI similarity searching that had been explained to them. Five users confessed to not understanding the filters. Although the others said they understood the concept of filtering the retrieved result set by industry sector, date of article, etc., none of

TABLE 6
Misunderstandings of metaphors or system functionality reported by users in de-briefing interviews

Misunderstanding	No. of users
Thesaurus structure and links	7
Browser controls	6
Similarity model	6
Filters	5
Bullseye results-browser metaphor	3
Query formulation	3
Abstract viewer	1

them actually used these functions. The encoding of relevance in the browser-bullseye metaphor was understood by most users, although the rationale for clusters was not clear for three subjects. Three subjects were not aware that they could enter their own keywords as well as picking them from the thesaurus, also two subjects were confused about how to enter Boolean operators in queries. The system did not support Boolean queries because these are incompatible with LSI searching which essentially operates a conjunction (AND) style search. On the positive side all users liked the results browser display, understood the link between the "hover text" summary and selecting articles to display in the abstract viewer, as well as finding query formulation easy.

In conclusion, although the usability evaluation gave the system a reasonable assessment, it did point to some reasons for poor recall in operating the thesaurus and, more importantly, poor user comprehension of the thesaurus structure, bullseye clusters and similarity based searching. However, when performance, usability and comprehension problems were examined at the individual level, no significant correlations were found (see Table 8). Furthermore, there were no obvious associations between particular types of user problem and performance, evident in the distribution of particular comprehension and usability problems (e.g. thesaurus/bullseye display) between better and worse performing subjects. System usage was sub-optimal by most subjects. Instead of using the search facilities effectively, as the behaviour analysis in the next section shows, a majority of the subjects browsed through the abstracts rather than using similarity searching or filtering results sets. The system metaphors fared reasonably well. The thesaurus suffered some design problems but the basic hierarchy of terms was accepted, and the bullseye metaphor for relevance encoding was easy to understand. Other reasons for poor performance, such as in sub-optimal behaviour and search strategies, were investigated by the behaviour analysis reported in the following section.

5. User Behaviour

Mental and physical user behaviour during the evaluation sessions were video taped. The video tapes were analysed by viewing the tapes to create a sequential record of behaviour categories using the following definitions:

Physical behaviours (*observed*)	
Type/edit query	data entry/edit text in query string area
Thesaurus query	double click on thesaurus term to enter it as a query
Execute search	send query to the database
Navigate thesaurus	using thesaurus controls to expand/contract the hierarchy display
Manipulate results	altering the number of results clusters, expand or move cluster
Change cluster categories	select clusters by industry or similarity
Navigate results	browse bullseye display inspecting "hover text" article summaries
Evaluate content	scan article abstract, scroll through article and document details
Find similar	use find similar articles control

Mental behaviours (*verbalized or observed*)	
Read article	read article details or abstract
Select terms	decide which term to use for a query
Select document	decide whether to select or reject a document.

Sequences were divided into segments and categorized with approximate timings to the nearest minute. The video tapes were replayed when sequences of behaviour were rapid and difficult to analyse. Two independent observers analysed selected segments of the sessions and their categorizations were compared. The initial inter-observer agreement was 77%. Differences were reconciled and common coding of behaviour categories was agreed. The frequency of behaviour by subject for both tasks is shown in Table 7.

Four behaviours—selecting relevant articles; evaluating abstracts; navigating results bullseye display by inspecting articles' titles; and reading abstracts—accounted for 85% of all the behaviour. Subjects were divided into better performers who achieved $> 10\%$ recall on merged data for both tasks and $> 5\%$ recall on each task, and worse performers. All of these behaviours, except reading the abstract and total behaviour frequency, were more frequent in the first six better performing subjects (JU to AG2 vs. RM to GS, $p = 0.05$ Binomial test comparing normalized scores for each group). Manipulation of the results (changing size or moving the bullseyes) and similarity searching showed considerable individual differences. To analyse patterns and search strategies, users' behaviours were scored in sequential order and cast into a matrix so that the frequencies of transitions between behaviours could be investigated (i.e. the number of times behaviour A was followed by behaviour B, A was followed by C, etc). The matrix was converted into a behaviour network diagram to illustrate the pattern of

TABLE 7

Frequencies of behaviours for each subject observed during both experimental tasks. Two behaviours, change cluster categories and select terms, occurred very rarely and were eliminated from the following analysis

Subject	Input query	Thesaurus query	Execute search	Navigate thesaurus	Manipulate results	Navigate results	Evaluate content	Find similar	Read article	Select document	Total
JU	1	5	4	2	0	79	77	0	0	40	208
AG1	3	15	5	5	10	62	56	2	17	38	214
HS	4	8	7	4	11	70	61	0	50	32	247
MK	5	14	6	6	7	51	36	4	7	34	170
KK	1	2	2	4	0	48	40	1	27	29	154
AG2	2	3	3	4	6	83	75	5	26	34	241
RM	1	3	3	2	1	36	28	4	4	19	101
RV	5	0	6	5	6	33	30	1	11	15	112
MD	3	4	2	3	5	25	23	0	11	16	92
AM	1	2	2	2	0	14	14	0	11	13	59
CN	0	2	2	2	0	8	8	0	8	6	36
GS	3	2	6	2	6	22	17	0	14	8	81
Mean	2.4	5	4	3.4	4.3	42.2	38.8	1.1	15.5	23.7	142.9

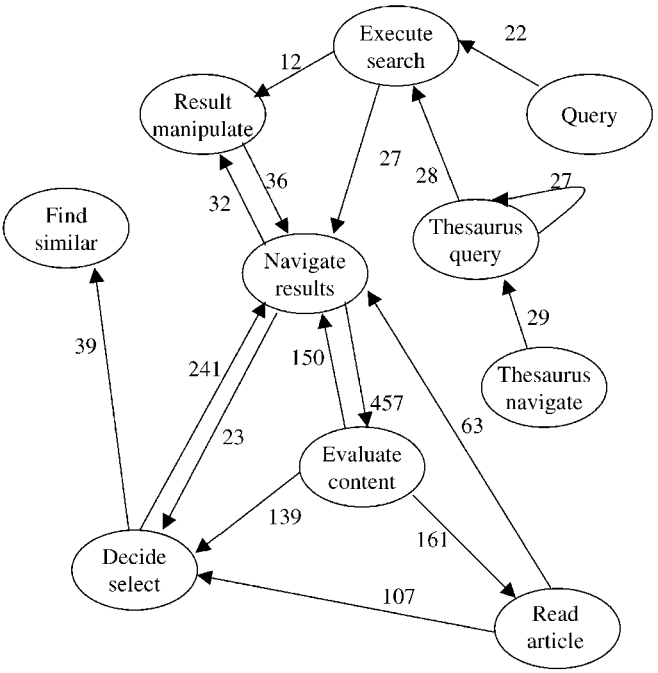

FIGURE 3. Behaviour pattern diagram for all subjects, illustrating only transitions > 1% of the total number of transitions.

behaviour during each session. The behaviour pattern for all subjects is illustrated in Figure 3.

The group-level pattern commenced with thesaurus navigation followed by selecting terms. Terms were directly entered as queries nearly as frequently as being selected from the thesaurus. Term selection led to query execution which took some time; consequently there were some breaks in the behaviour sequence at this point. At first users tried to continue interaction but they quickly learned that the system did not support multi-tasking and patiently waited until the results were displayed. The most frequent pattern was then to browse through the bullseye results display, although some users manipulated the display (usually enlarging and moving the circles) before browsing. The evaluation cycle, which started with browsing the bullseye display, was followed by evaluating the abstract by scanning, reading it through and then deciding whether to select or reject it as relevant to the task. This concluded the task in most sequences; however, in a minority of sequences the users proceeded to use the similarity search function. This was followed by more article evaluation or search termination, but it is not shown on the diagram as these transitions fall below the 1% cut off. The absence of a behaviour cycle linking the thesaurus with results evaluation or document selection, combined with the low number of queries submitted by most subjects, indicates that the users did not refine queries; instead they concentrated on evaluation of retrieved articles for a small number of queries.

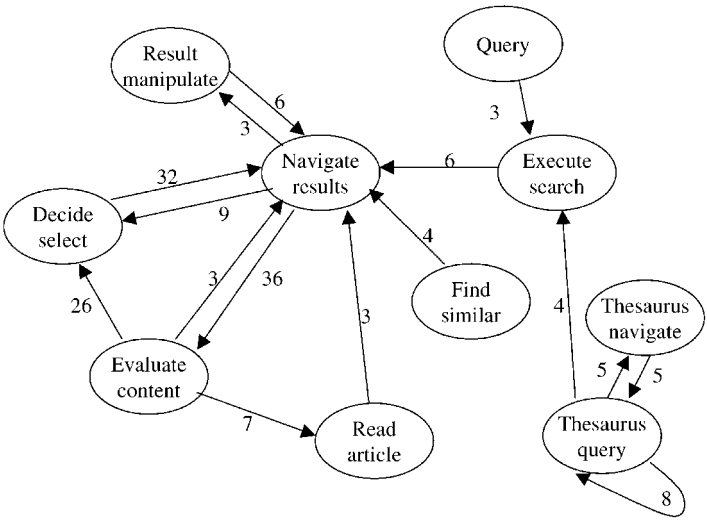

FIGURE 4. Behaviour pattern for subject MK, who was the 4th best performer in % recall. Transitions $< 1\%$ total for this subject have been omitted. This subject shows a rich behaviour pattern similar to the merged view of the whole group.

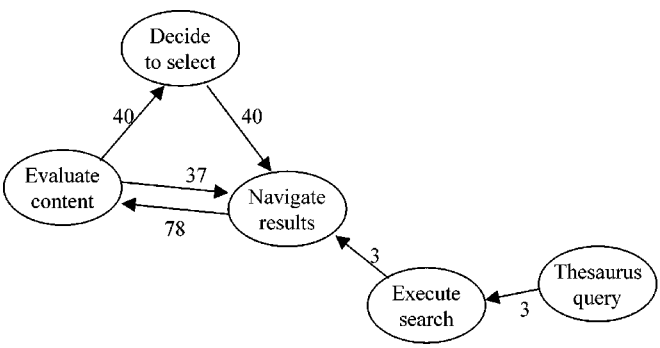

FIGURE 5. An impoverished behaviour pattern of subject JU (best performer) who did not explore the thesaurus beyond picking keywords and showed little manipulation of the results display. However, this subject did show a high frequency of transitions between Navigate results and Evaluate content, not observed for other impoverished pattern subjects.

Individual patterns either followed the group level with minor variations (see Figure 4) or showed an impoverished pattern with no manipulation of the results browser display or use of the thesaurus, as illustrated in Figure 5.

The behaviour patterns show some association with performance measures, but the picture is not consistent. Five out of the six top performers had rich behaviour patterns, but subject JU was the exception, achieving an average 16.7% recall for both tasks with an impoverished pattern (see Figure 5). However, JU did show a high frequency of transitions between navigating results and evaluating content, reflecting a systematic approach to assessing the relevance of articles. Three out of the six poorly performing

subjects ($< 5\%$ recall in at least one task) showed impoverished patterns (RV, AM, CN). The other three had richer patterns but their frequency of, and transitions between, navigating results and evaluating content were low (Table 8). These subjects read the article title on the bullseye display and made decisions to select or reject a document without viewing the abstract, so it appears that a successful strategy may require a richer pattern with frequent transitions between exploring and evaluating the results.

The association between performance, errors and search behaviour is summarized in Table 8. As noted earlier, there were no correlations between error frequencies and performance. This is not surprising as the error data contain missing user requirements as well as usability problems; furthermore, the users did not experience severe usability problems that prevented task completion.

There was a tendency for the six better-performing subjects to actively explore and navigate both the thesaurus and results browser visualizations, whereas the poorer performers only navigated the results, and did so less frequently than better performers, apart from the two subjects (RV and RM) who were the top two in the worst performer group. There is a slight tendency for better performers to submit more queries but this was not significant. Subject GS submitted many queries for no reward, although his performance may have been impaired by poor motivation, as he had to be encouraged by the experimenter to keep going several times during the task and he selected very few articles (total of nine for both tasks). Indeed, four of the six worst-performing subjects complained that they found assessing document relevance was difficult, so a combination of poor motivation and lack of domain knowledge may have accounted for their poor performance. Time spent on the task does show a positive association with performance, although this failed to achieve significance (Spearman rank order correlation coefficient). Longer task completion times, observed for three of the six better-performing subjects, were associated with more transitions between navigating results and evaluating retrieved articles rather than by other behaviours. In contrast, subject JU spent a short time on both tasks, but compensated for this by evaluating many articles carefully and selecting a large number, so there appear to be two explanations for good performance: longer task completion times and exploration of both thesaurus and results browser displays; and three explanations for poor performance: shorter task completion times, poor use of both the thesaurus and the results browser displays and possibly poor motivation leading to impaired evaluation of articles' relevance.

In conclusion, it appears our visualization was at least partially successful. A correlation between post-test misunderstandings and performance may have been expected but the data did not support this hunch, so the subjects could achieve good results in spite of sub-optimal system usage. The upside of this is to conclude that the design was robust, while the downside is that poor understanding may have restricted our subjects' retrieval performance. Usability problems did not seem to inhibit effective operations of the system. However, system facilities usage was sub-optimal and this may have constrained users from achieving better performances. Many of the problems we discovered with the users' poor understanding of the system model probably impaired performance in absolute terms, even if this did not appear to contribute to explanation of individual performance differences.

TABLE 8

Summary of user performance, usability problems, and search behaviour. The strategies are taken from the behaviour analysis, reporting only active navigation of the thesaurus and results visualizations. Recall is % for both tasks; time = Task completion time in h.mm. Infrequent navigation was scored when fewer than 20 transitions were observed between Navigate results and Evaluate content; other subjects scored > 40 transitions

Subject	Recall	Usability problems in tasks	Problems reported post-test	Report misunder-tandings	Searches submitted	Time	Strategies and behaviour
JU	16.7	5	4	1	4	0.18	Frequent navigate results
AG1	16.7	2	6	4	5	1.15	Navigate thes + results
HS	14.3	0	6	2	7	0.42	Navigate thes + results
MK	14.3	7	9	4	6	0.29	Navigate thes + results, similarity search
KK	14.3	8	7	0	2	0.34	Navigate thes + results
AG2	11.9	5	6	1	3	1.05	Navigate thes + results, similarity search
RM	9.5	7	5	6	3	0.22	Navigate results
RV	9.5	6	5	4	6	0.40	Navigate thes + results
MD	7.1	8	6	2	2	0.23	Infrequent navigate results
AM	7.1	1	4	0	2	0.21	Infrequent navigate results
CN	4.8	6	6	3	2	0.40	Infrequent navigate results
GS	2.4	7	4	4	6	0.26	Infrequent navigate results

6. Discussion

Although our visualization design appeared to pass the usability test apart from some minor glitches, user performance was poor. One reason for this may have been the relative unfamiliarity of the system. When users did interact with the visualizations, better results were achieved, so one lesson may be that more training is required. The post-test debriefing showed that the subjects had only partially understood the system metaphors and functionality, e.g. the clusters and filters. Furthermore, most of our subjects used the system with conservative strategies consisting of simple queries without cycles of refining searches. This contrasts with search behaviour that we, and others, have observed with traditional information retrieval interfaces (e.g. MEDLINE with WinSpirs) in which expert users refine queries using narrowing and broadening strategies (Marchionini, 1995; Sutcliffe, Ennis & Watkinson, in press). It is possible that visual browser interfaces inhibit such behaviour. As information search tools will often be used by end users with little training, for instance in WWW applications, expert assistants (i.e. wizards and guided tours) may be necessary to explain more complicated visualizations and how to use them effectively with efficient information searching strategies.

The implications for visualization designers from our study are that a more systematic approach to developing appropriate combinations of visualization and functionality is needed. Our design was motivated from a task model of information searching and a data model for the thesaurus and results browser. The visualization did appear to be comprehensible to users; however, it was hindered by lack of guidance on search strategies and possibly by the manipulations we provided for exploring the thesaurus and results browser visualizations. Basing visualization design on user tasks and data models has been advocated by others (Card *et al*., 1999) and demonstrated in successful products (Ahlberg & Shneiderman, 1994); however, in more complex tasks further research on visualization design methods that integrate active system guidance with visual browser and exploration tools is required.

Our previous studies on information retrieval showed that user strategies were one important determinant of search success (Sutcliffe *et al*., in press) as have other studies (Kulthau, 1993; Marchionini, 1995). Another success factor noted in several studies is choice of appropriate search terms (Marchionini, 1995; Ingwersen, 1996; Sutcliffe *et al*., in press). We provided a visual thesaurus to tackle this problem; however, several users commented that the terms in the thesaurus did not match their expectations and typed in their own queries. One lesson here is that visual structure is no substitute for either a well-designed thesaurus or user-customizable thesauri. We had included a user customization facility in our design but did not test it because of increasing the complexity of an already complex system for novice users. The mismatch between terms, classification of terms and visualization structure remains a subject for future research.

The representation of results sets with the bullseye metaphor was successful, and encoding both relevance and similarity was understood by our users; hence the bullseye display appears to be an advance on simple display boxes for representing similar categories as in Scatter/gather (Pirolli *et al*., 1996). However, closer examination of the post-test interviews showed that while the users understood the relevance ranking metaphor, their concept of functionality by which similarity was calculated by the search process was less clear. This may have been caused partly by the users' current mental

model of Web search engines overriding understanding the LSI algorithm (most users thought similarity was just shared keyword frequencies). However, LSI can occasionally rate dissimilar documents as being similar, particularly when it is being used for matching between a small number of keywords and a whole document. In our system, LSI occasionally produced some results which did not seem to belong in a particular cluster, consequently confusing the users. Another reason is probably lack of training. Our users did not make extensive use of the similarity search function although a majority stated that they did understand it.

The behaviour analysis produced a bimodal distribution between the more adventurous subjects who explored with visualization and those who were more conservative and just picked keywords and selected results. Users' prior experience did not correlate with this grouping, neither did gender, and we did not screen pre-test with a visualizer–verbalizer cognitive inventory (Leutner & Plass, 1998), so we have no explanation beyond individual difference for this effect. However, search success was not directly attributable to the users' patterns of interaction alone; instead users had to spend considerable time carefully evaluating articles as well as using the system visualization effectively. Although the bullseye results browser was approved of and used by all the subjects, it did not help the poor performers. Summarization of results can improve user performance compared to simple relevance ranking (Pirolli *et al*., 1996), and our subjects preferred the results browser to Web search engine ranked lists; however, we found that there are considerable individual differences in people's effective use of such visualizations. Visualization tools need to encourage users to carefully inspect document contents, possibly by marking keywords in documents or hit density maps, as well as presenting overview summaries.

In our future designs we will change the tiled window screen layout, considering the complaints about the representation of the thesaurus. Also our users did not refine queries in iterative use of results browsing, thesaurus navigation and querying, so the rationale for concurrent visualization of all facilities relevant to the task was not supported. The user errors and misunderstandings suggest that a larger thesaurus display with customization facilities would be preferable. Another problem was the tendency of the visualization to bias users away from selecting their own keywords. More positively, the results browser visualization did work well with the document viewer, and these tools helped users to assess article relevance from titles on the bullseye display in combination with the document viewer. However, some of the poorly performing subjects assessed articles only from the bullseye display, so this illustrates a potential hazard of visualization tools encouraging sub-optimal and cognitive lazy practice.

The success of the visual metaphors in explaining system functionality and representation of data to users was mixed. The relevance ranking and cluster-similarity grouping metaphors were understood; however, user debriefing demonstrated that there was considerable confusion about the identity of groups of documents which the system had rated as being similar. This problem, also encountered in Scatter/gather (Pirolli *et al*., 1996), limited the system's effectiveness because the users found it difficult to relate the clusters to their query. Similarity clustering of results may therefore not be helpful unless it is directly related to the user's query and clusters are labelled with terms related to the query. However, this requires considerable inference to analyse shared properties of all

the documents in a cluster, so most systems have to rely on manual labelling. Furthermore, the clusters have to be cohesive and consistent with the users' view of a "logical group".

Generally, our system was under-utilized. The reasons for this seem to be a combination of usability problems which impaired the effectiveness of the thesaurus, and inadequate training for more complex features such as similarity based search. In spite of under-utilization, poor performance was not directly attributable to the visualization design. Unfortunately, good visual design may be no panacea for poor search performance attributable to user motivation and lack of domain knowledge. While some studies have shown that visual information retrieval tools can improve performance over simple ranked list displays (Chen & Dumais, 2000), this study has raised a cautionary note about individual differences and demonstrates that improving users' search performance may require training and system assistance in search strategies and assessing relevance that is integrated with visual representations of meta-data and retrieved documents.

This research was partially supported by ESRC Cognitive Engineering project MISSAR (Modelling Information Seeking Strategies And Resources) and EU Telematics project Multimedia Broker.

References

AHLBERG, C. & SHNEIDERMAN B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays. In B. ADELSON, S. DUMAIS & J. OLSON Eds. *Celebrating Interdependence*: *CHI '94 conference proceedings on Human Factors in Computing Systems*, pp. 313–317. Boston. New York: ACM.

CARD, S. K., ROBERTSON, G. & MACKINLAY, J. D. (1991). The information visualizer: an information workspace. *Proceedings of CHI'91*, *Conference on Human Factors in Computing Systems*, pp. 181–188. New Orleans. New York: ACM.

CARD, S. K., MACKINLAY, J. D. & SHNEIDERMAN, B. (1999). Information visualization. In S. K. CARO & B. SHNEIDERMAN, Eds. *Readings in Information Visualization*: *Using vision to Think*, pp. 1–34. Los Altos, CA: Morgan Kaufmann.

CHEN, H. & DUMAIS S. T. (2000). Bringing order to the Web: automatically categorising search results. *Proceedings of CHI '00*, *Conference on Human Factors in Computing Systems*, Den Haag, April. New York: ACM.

CZERWINSKI, M., VAN DANTZICH, M., ROBERTSON, G. G. & HOFFMAN, H. (1999). The contribution of thumbnail image, mouse-over text and spatial location memory to Web page retrieval in 3D. *Proceedings of Interact '99*, Edinburgh, 1–4 September.

HANCOCK-BEAULIEU, M., FIELDHOUSE, M. & DO, T. (1995). An evaluation of interactive query expansion in an on-line library catalogue with a graphical user interface. *Journal of Documentation*, **51,** 225-243.

KULTHAU, C. (1993). *Seeking Meaning*. New York: Ablex.

INGWERSEN, P. (1996). Cognitive perspectives of information retrieval: interaction elements of a cognitive IR theory. *Journal of Documentation*, **52,** 3–50.

LANDAUER, T. K. & DUMAIS, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, **104,** 211–240.

LEUTNER, D. & PLASS, J. L. (1998). Measuring learning styles with questionnaires versus direct observation of preferential choice behaviour in authentic learning situations: the visualizer/verbalizer behavior observation scale (VV-BOS). *Computers in Human Behavior*, **14,** 543.

MARCHIONINI, G. (1995). *Information Seeking in Electronic Environments*. Cambridge: Cambridge University Press.

MONK, A. G. & WRIGHT, P. (1993). *Improving your Human–Computer Interface*: *A Practical Technique*. Englewood Cliffs, NJ: Prentice Hall.

PIROLLI, P., SCHANK, P., HEARST, M. & DIEHL, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver. New York: ACM.

PLAISANT, C., MILASH, B., ROSE, A., WIDOFF, S. & SHNEIDERMAN, B. (1996) Lifelines, visualising personal histories. *Human Factors in Computing Systems*: *Proceedings of CHI 96*, Vancouver, pp. 221–227. New York: ACM.

SUTCLIFFE A. G. & ENNIS M. (1998) Towards a cognitive theory of information retrieval. *Interacting with Computers*, **10,** 323–351.

SUTCLIFFE A. G, ENNIS, M. & WATKINSON, S. J. Empirical studies of end-user information searching. *Journal of the American Society for Information Science* (in press).

SUTCLIFFE, A. G. & PATEL, U. (1996). 3D or not 3D: is it nobler in the mind? In M. A. SASSE, R. J. CUNNINGHAM & R. WINDER, Eds. *People and Computers XI*: *Proceedings of HCI-96*, London, pp. 79–93. Berlin, Springer Verlag.

SUTCLIFFE, A. G., RYAN, M., SPRINGETT, M. V. & DOUBLEDAY, A. (2000). Model mismatch analysis: towards a deeper explanation of users' usability problems. *Behaviour and Information Technology*, **19,** 43–55.