Testing Visual Information Retrieval Methodologies Case Study: Comparative Analysis of Textual, Icon, Graphical, and "Spring" Displays

Emile Morse* and Michael Lewis
School of Information Sciences, University of Pittsburgh, 135 N. Bellefield St., Pittsburgh, PA 15260.
E-mail: emile.morse@nist.gov; ml@sis.pitt.edu

Kai A. Olsen
Molde College, Norway, N-6401 Molde. E-mail: kai.olsen@hiMolde.no

Although many different visual information retrieval systems have been proposed, few have been tested, and where testing has been performed, results were often inconclusive. Further, there is very little evidence of benchmarking systems against a common standard. An approach for testing novel interfaces is proposed that uses bottom-up, stepwise testing to allow evaluation of a visualization, itself, rather than restricting evaluation to the system instantiating it. This approach not only makes it easier to control variables, but the tests are also easier to perform. The methodology will be presented through a case study, where a new visualization technique is compared to more traditional ways of presenting data.

Introduction

Modern technology has given us the ability to store, retrieve, and disseminate huge amounts of documents. At the same time, our ability to read the documents returned from our searches to the Web or the bibliographic databases is as slow as ever. Where the retrieval systems may return hundreds to thousands of documents, we may only study a pitiful few.

Visualization techniques appear promising as a means for overcoming this problem. By mapping contents and relevance onto multidimensional spaces the user may be given an overview of a document collection, helping her to retrieve the documents of interest. In the last decade close to a hundred different visualization systems have been proposed. Very few of these have been subjected to user studies. One reason for this may be that full-featured systems are complex to evaluate. It is difficult to choose a control system, performance is related not only to the task but also to the mode of interaction, the computer resources, and to the choice of features. Other problems are the number of prototype systems that exist, and the dynamic nature of most systems and the costs of user training.

In this article we propose, through a case study, a basic stepwise (BASSTEP) method for testing such interfaces. Instead of putting the full system on the bench, we concentrate on the basic visualization strategy behind the systems. This allows for simple test situations, where it is possible to run many subjects, thus providing robust results, which can provide a scientific base for further development of visual systems.

A set of basic displays will be used to determine the effectiveness of different techniques. The tests presented here start with simple two-term Boolean situations, then move to three terms and finally, vector representations of two and three terms. Tests were performed both on paper and over the Web.

To determine the robustness of the results the Boolean test cases have been run in two different settings, using subjects from two different cultures (United States and Norway). All together more than 600 subjects took part in the study (216 for the two-term Boolean tests, 223 for the three-term Boolean, 196 for the more advanced vector techniques).

Visual interfaces for information retrieval and browsing can take many forms including visualizations based on reference points (Korfhage, 1991; Korfhage & Olsen, 1997), maps (Lin, 1991), or hierarchical graphs (Hearst & Karadi, 1997). Several of the more common visualization strategies are representing multidimensional values of a

* Present address: National Institute of Standards and Technology, 100 Bureay Drive, Stop 8940, Gaithersburg, MD 20899.

Received January 5, 2000; Revised May 8, 2001; accepted May 24, 2001

DOI: 10.1002/asi.10006

TABLE 1. Reference point visualizations.

Visualization type	System
Word	Ordered text such as search engine output
Icon list	TileBars (Hearst, 1995), Cougar (Hearst, 1994), Veerasamy & Belkin, 1996; Veersamy & Heikes, 1997, NIRVE (Sebrechts et al., 1999), SFA (Shaw et al., 1999) Kakimoto & Kambayashi, 1999
Graph (Cartesian)	GUIDO (Nuchprayoon, 1996), BIRD (Kim & Korfhage, 1994), InfoCrystal (Spoerri, 1993), Component State Drawing (Crouch & Korfhage, 1990), DARE (Zhang & Korfhage, 1999), Bead (Chalmers & Chitson, 1992), SENTINEL (Fox et al., 1999), SPIRE (Wise, 1999), NIRVE (Sebrechts et al., 1999), Asp-Inquery (Swan & Allan, 1998), SFA (Shaw et al., 1999), TETRALOGIE (Mothe & Dkaki, 1998), Kakimoto & Kambayashi, 1999
Spring (physical analog)	VIBE (Olsen et al., 1992), Radviz (Au et al., 2000)

document as attributes of a glyph, positioning document glyphs according to dimensional weighting in a 2D or 3D Cartesian space, or positioning document glyphs by equating dissimilarity and distance. This last strategy can vary along a continuum from a single reference point serving as the graphical origin to map displays that organize the entire space taking into account all interdocument similarities (Lin, 1996). Implemented systems commonly follow Mackinlay's (1986) approach of using multiple visualization strategies to convey additional dimensions. SFA (Shaw et al., 1999), for example, displays six dimensions by mapping three to glyph location and three to the glyph attributes, size, shape, and color. Other approaches such as spring embedding (Chalmers & Chitson, 1992), Kohonen maps (Lin, 1991), and nonorthogonal "axes" (Au, Casey, Sewraz, Guo, & Ruger. 2000; Olsen, Williams, Socharts, & Hirtle, 1992) seek to approximate high dimensional distances in 2D or 3D displays.

The present study restricts itself to four visualization strategies (Table 1) commonly used in reference point based visualizations although the BASSTEP methodology can be applied to evaluating any visualization strategy that can be clearly characterized.

Table 1 presents a representative set of reference point systems. Each of these relies on a representation of documents as vectors, although systems such as InfoCrystal contain Boolean vectors. Several of these systems along with visualizations primarily for Boolean values will be described to illustrate the range of implementations employing these visualization strategies and the difficulties intact systems present to evaluation.

Component Scale Drawing (Crouch & Korfhage, 1990) uses a graph that shows query terms on the x-axis; the order of the terms is determined by the user's weighting. The y-axis indicates classes of term weights. Documents are represented as broken lines, and the query itself is represented as a solid line. The purpose of the system is to assist users in determining the similarity of query and documents.

Cougar (Hearst, 1994) uses a Venn diagram to represent the relationship between documents and query terms. Each query term is mapped to a circular area of the display. Document identifiers are shown as icons in the list box in the appropriate sector of the graph. GUIDO (Nuchprayoon, 1996) uses a novel type of display to allow sophisticated mathematical manipulation of similarity metrics. The display allows the selection of two reference points that are then shown as points on the x- and y-axes. The resulting document set is then displayed in the plank that is generated at a 45° angle in the graph. Various retrieval caps and metrics are provided to enhance selection of desirable subsets of documents.

InfoCrystal (Spoerri, 1993) is another example of a reference point system that is based on the Venn diagram model. Figure 1 shows the results of a four-term Boolean query. The query terms are indicated at the vertices, and the resultant subsets are associated with the other shapes shown in the bounding box. The number of edges of an included shape indicates the number of query terms and the direction of the vertex points to the query term. InfoCrystal is useful for determining the distribution of documents in a document collection that satisfy each of the possible Boolean queries. Spoerri describes higher dimensional InfoCrystals. He also illustrates a version that allows the creation of weighted vector queries, although the display looks tremendously complex.

SIRRA (Aalbersberg, 1995) incorporates a list of multicolor icons. Each query term is assigned a color, and each icon represents a single document. Users can compare documents with respect to the strength of a query term within

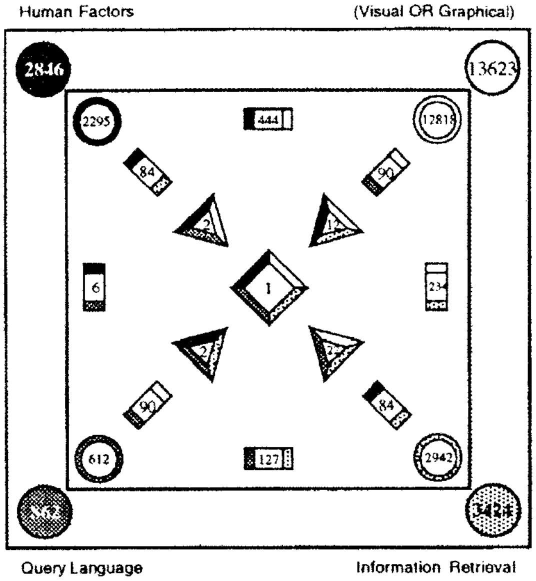


FIG. 1. InfoCrystal.

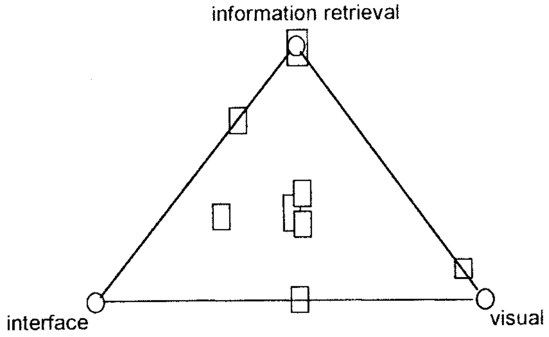


FIG. 2. Vibe display.

and across documents in a set. Space (Newby, 1992) is an IR system based on the principles of navigation, which he defines as "human behavior to make sense of an information space." In the part of the interface termed the "Navigation window," keywords and document identifiers float in the field. The placement of the documents with respect to the keyterms is based on the relative strength of attraction of the document for the term. Other windows in the interface include a map view and a key term list. TileBars (Hearst, 1995) is based on segmentation of the underlying full-text into topics, where grayscale rectangular areas show the relative amount of the term in sequential fragments of the text.

An example of VIBE (Olsen et al 1992) is shown in Figure 2. Query terms, called points of interest (POIs), are shown at the vertices of the figure. A POI may be as simple as a set of keywords, or may be defined by more complex functions. Documents receive a score on each POI, for example, by a frequency count on the POI terms. The POIs define an information space, and document icons are located in this space according to the score, for example, in the middle of the POIs with an even score on each POI, on top of a POI if all other POIs give a zero score, etc. VIBE can be used to represent the results of Boolean as well as vector queries. This visual is part of a fully featured interface that allows users to interact with multidimensional displays (i.e., any number of POIs) and moderately large document collections.

Several types of visual approaches can be seen when the above examples are analyzed. Four major categories are: Venn diagram, icon lists, and two types of spatial systems. Both Cougar and InfoCrystal are based on the Venn diagram. SIRRA and TileBars present icon lists. VIBE and Space employ a spatial method to render the relationship between documents and key terms. GUIDO and Component Scale Drawing provide graphical representations based on an x-y graph. Component Scale Drawing uses a line graph and nominal scales while the GUIDO uses icons and continuous scales.

The prototypes described for testing in the present studies contain representatives of the icon list and spatial types of displays. The Venn diagram appears to be useful for displaying Boolean data, but falls short of making compelling displays of vector data. The x-y graph type that is to be used in the two-term prototype testing is similar to GUIDO and Component Scale Drawing and closely related to another IR visualization, BIRD (Kim & Korfhage, 1994).

Evaluation of Visual Interfaces for Information Retrieval

Of the reference-point visualizations introduced, only Component Scale Drawing, GUIDO, Space, TileBars, VIBE, AspInquery, NIRVE, and the Veerasamy visualization tool have been subjected to controlled user studies. Each of the studies will be reviewed briefly here. The purpose in reviewing these evaluation methods is to determine what tasks were given to the subjects, and also to determine which interfaces were subjected to usability evaluation as opposed to task performance evaluations. Other pertinent aspects of the studies, such as number of subjects used and characteristics of the user/subject populations, will be noted where such information exists.

Component Scale Drawing (Crouch & Korfhage, 1990) was tested by presenting a user with a display based on each of 15 queries. The task of the user was to apply the Component Scale Drawing tools to rank the documents with respect to their similarity to the underlying query. The rankings were then compared with the known relevance rankings. The results showed that there was a highly significant relationship between the user's rankings and the known rankings (Spearman coefficient .85 across queries). The number of users is not clear from the article, but may be limited to a single person. The task is clearly highly specific to the interface.

GUIDO was subjected to usability testing (Nuchprayoon, 1996). Sixteen subjects were charged with performing nine information retrieval tasks. Tasks were graded as "easy" and "hard"; "easy" tasks presented the test subject with preselected metrics, retrieval threshold, and POIs, while "hard" tasks required the subject to select each on his own. The primary goal of each task was to choose the eight "best" documents from the resultant display. The primary measure was the amount of time that it took the subjects to perform the document selection. The results showed that there were some interactions between the retrieval threshold and metric. Subjects provided positive feedback on the GUIDO system.

Newby (1992) tested Space with 20 users. They were provided with a full system display that included the multiwindow display and a mouse and PowerGlove. His primary goal was to test the ability of users to navigate abstract spaces. Users performed two information retrieval tasks: (1) a closed-ended question that was based on key-term synonymy, and (2) an open-ended task based on a vague statement of information need. The "Space" system was compared with a traditional IR system (Prism). Newby demonstrated considerable learnability of the Space system and high user ratings. Comparison with the more traditional system showed that users preferred the system with which they were already familiar.

TileBars has not been subjected to the same type of user studies mentioned thus far. The TileBars interface itself has not been user tested, but the algorithm underlying its segmentation of text into topics has been compared with human segmentation of the same text. High correlations were found between the two types of segment generators (Hearst, 1995). This study, however, is not particularly relevant for the purposes of the proposed work.

VIBE has been subjected to user testing by Koshman (1996). She compared performance of expert and novice searchers using VIBE or a conventional text-based interface (AskSAM). There were 15 novices, 12 on-line search experts, and four subjects who had VIBE system expertise. Due to the small sample of VIBE experts, the study concentrates on the former two groups. This was a thorough usability study of the VIBE interface in that it sought to measure users' performance at tasks that required use of novel interface features. Subjects performed seven tasks that were chosen for their likelihood to represent "normal" user IR tasks. In general, the tasks have a Boolean flavor, for example, how many documents contain (all, one, or two) terms. Scenarios were constructed to provide a naturalistic information seeking setting. Usability was assessed by measuring: (1) system familiarity time, (2) task performance speed, (3) frequency with which on-line help is accessed, (4) number of errors in task results, (5) subjective satisfaction, and (6) system feature retention. Familiarity time showed no difference for interface or for expertise level. She showed that time to complete tasks was inversely related to expertise. Users preferred the familiar, text-based interface to the visual VIBE interface. She states that users retained what they learned from one session to the next but believes that this was due to increased "familiarity with the kinds of tasks and the tools needed to perform the tasks." It is reasonable to conclude that the Boolean nature of the tasks chosen for this study influenced the outcome, in that Boolean tasks are probably accomplished more effectively with Boolean systems such as AskSAM.

Swan and Allan (1998) compared a control text-based retrieval system ZPRISE, AspInquery, an added "aspect window" intended to help users organize information they have retrieved, and AspInquery Plus which added a 3D spring-embedding visualization of the retrieved documents. The display added in the AspInquery Plus condition is the study's spatial visualization. Twenty-four users evenly divided between university librarians (experts) and recruited students (novices) participated. Participants performed six topic searches on topics selected by NIST from previous TREC experiments and measurements were taken for recall, precision, and time. As in Koshman's study the librarians overwhelmingly preferred the textual interface; however, the students preferred the visual interfaces. No performance advantages were found for the spatial visualization or expertise.

NIRVE uses a hybrid visualization combining aspects of the icon list, spatial clustering, and graph representations making inferences about particular visualization strategies difficult. Sebrechts, Vasilakis, Miller, Cugini, and Laskowski (1999) compared a textual control condition with 3D and 2D versions of the NIRVE display in an experiment with 15 participants. Eight groupings of task types were identified for testing. Once again the textual control produced the fastest response times. Training improvements were greatest in the 3D condition, and for participants with high self-rated computer skills response times for 3D and textual interfaces were comparable.

Veerasamy has conducted a series of experiments investigating an icon list-type visualization tool. In two initial experiments reported in Veerasamy and Belkin (1996) TREC-4 interactive track topics were used as tasks. In the first experiment 36 subjects were divided into groups getting the visualization, no visualization, or visualization only on their second topic. Twenty-four of the 25 topics were used with two assigned per subject. No differences were found in precision, documents saved per search, interactive TREC precision, or interactive user precision. In the second experiment, variance due to search topic was controlled by assigning the same three topics to 36 subjects. In this experiment one of the three topics showed improved performance for participants in the visualization condition. The next year Veerasamy and Heikes (1997) were able to demonstrate clear performance advantages of their visualization tool by extending control even further. In the new experiment, the researchers focused on relevance judgments, a task more directly tied to the visualized information, and one in which the 37 participants were required to make precisely the same decisions. Under these conditions subjects using the visualization tool were both more accurate and rapid in their relevance judgments.

The most striking thing about this research survey (Table 2) is how infrequently information visualizations have been user tested and how poorly they fared when they were. We believe this is due to a number of factors:

(1) Intact systems are too difficult and complicated to learn for users to become adequately skilled over the course of an experiment. Morse and Lewis (1997) showed that a drastically simplified version of VIBE produced large gains in usability. The training effects reported by Koshman (1996), Newby (1992), Swan and Allan (1998), and Sebrechts et al. (1999) point in this direction as well.
(2) Tests need to be more tightly controlled. This is especially true of open-ended tasks where participants may be doing very different things. Controlling tasks and identifying tasks dependent on information supplied by a visualization can greatly improve the likelihood of finding effects.

The BASSTEP methodology we are proposing provides a systematic approach for dealing with both the training and control issues.

TABLE 2. User studies of reference point based visualizations.

Study	N	Task	Visualization effect	Preference	Text control
CGD	—	Relevance judgment	$R = 0.85$	—	No
GUIDO	16	Representative tasks	—	—	No
Space	20	1 restricted 1 open ended	No difference	Text based	Yes
TileBars	—	Correlate system/human text segmentation	—	—	No
VIBE	31	Representative tasks	No difference	Text based	Yes
AspInquery Plus	24	TREC topic searches	No difference	Text based	Yes
NIRVE	15	Representative tasks	Text favored	—	Yes
Veerasamy, 1996	72	TREC topic searches	No difference	—	Yes
Veerasamy, 1997	37	Relevance judgments	Visualization favored for time and accuracy	—	Yes

BASSTEP Methodology

User testing of a computer system is not a simple task. Full systems include a large set of variables that may be difficult to control, for example:

(1) *Computer experience*. The subjects may have varying degree of computer experience, from expert programmers to subjects that do not feel comfortable with a mouse or keyboard.

(2) *Look and feel*. The subjects may have experience with programs that have the same look and feel as the test system, or their background may be from quite different platforms.

(3) *Computer equipment*. The equipment used for the studies may influence the results. For example, it seems natural to believe that screen size and resolution may have an impact on studies of visualization.

(4) *Training*. Testing a full system will often require a training session that may influence test scores. In some cases training may be viewed as an integral part of using the system; however, training subjects for a user study may often be different from training in a real-life situation.

(5) *Functionality*. A full system will often include a set of functions that supports the basic visualization method, for example, search tools, filters, zoom and pan features, open and save operations, print options, etc., but which are not a part of the visualization methodology itself. The application of such tools may influence test results.

(6) *Time of running tests*. Tests on a full system will be time-consuming, especially if training is involved. This makes it more difficult to recruit subjects over the full range of potential system users.

(7) *Cost of running tests*. The need for labs, training, and long sessions make testing expensive. Thus, we find that most user testing of full systems is performed with a very limited number of subjects, perhaps as low as 10. Then it becomes very difficult to control the influence of context variables, especially background and training.

Novel visualization techniques are most often presented in the form of a prototype system. Using complete prototype systems for user testing may introduce additional complications. Prototype systems may be unstable, may only work on one platform, and will not have a mature set of supporting functions. Because such systems are in a development phase, and because the available features may influence results, results from previous versions may not be applicable to current or future versions. Thus, as we recognize the importance of user studies of novel visualization techniques, we find that it is extremely difficult to get reliable results from studies of full systems.

The important question is then: is it really necessary to test the system itself? The answer is "yes" when our aim is to evaluate different implementations of the same methodology. For other situations, where we are discussing the merit of a visualization methodology, the methodology itself may be studied apart from the implementation. To show how this can be done, the VIBE visualization methodology shall be used as a case study. VIBE has been implemented in both research prototypes and commercial programs, in a 2D and a 3D space, directed towards record-based, document, or image data. VIBE implementations have been developed on many platforms (Unix/X-window, Windows/C-program, Windows/Visual Basic/SQL database, Web/Java Applet), each with a large set of supporting functions. All of these implementations use the same basic visualization methodology that was presented earlier, an information space defined by POIs (point of interest), a score for each object on each POI, and displaying the object icon with regard to the POIs. The basic idea behind this display is that the position of an icon should *intuitively* give clues as to the contents of the object (record/document/image) with regard to the POI.

It turns out that a study of the BASIC ideas is much easier to perform than a study on a full system. For example, a test to see if subjects can "read" a VIBE display can be performed on paper. Because a VIBE display should be intuitive, no training will be needed. Note also that we have effectively removed all influence from computer experience, "look and feel," equipment, etc. Paper tests like these are so inexpensive and simple to run that they can be tried on large numbers of subjects. In the case studies that are presented in the next section, several hundred subjects were used in each study. These tests may be even more efficient if moved to the Web, allowing users to enter response data directly. Note that this is still a simple test of the visualization basics, and can be performed by anyone with a minimum of knowledge of how to use a browser. By keeping example displays within the resolution limits offered by standard displays, effects of different types of display equipment can be controlled.

Paper (or a similar test over the Web) may be used for testing the static features of visualization techniques, and to some degree, also dynamic features (e.g., by presenting sequences of static displays). However, where the interactive and dynamic features of a visualization technique are to be studied these tests may be too limited. However, instead of testing the full system in this next step, we have good experience by using a *defeatured* system, i.e., a system that only includes the basic features that are to be studied (Morse & Lewis, 1997). We have implemented such systems either by removing all but the basic functionality of existing system, or by building new prototypes that only address the most basic features. Either way, we achieve simple, clean systems that are stable, easy to learn and to use, and which emphasize the underlying visualization methodology. For VIBE, these BASIC systems have all the important display functionality, presenting a set of data in a user-defined information space, allowing the user to dynamically reposition points of interest, clicking on icons to view documents, etc.

The overall aim of these studies was to determine whether defeaturing existing IR interfaces could produce interfaces that could be used successfully in "walk-up" systems, especially on the Web. The results showed that users could indeed form correct inferences about retrieved documents and their relationship to the query terms without extensive training when the VIBE interface was simplified.

Thus, with this BASIC methodology we are able to increase confidence levels by (1) focusing on the display methodology; (2) reducing the influence from context variables; and (3) using larger number of subjects, and at the same time make it easier to make generalized statements based on the results. For example, results obtained from a BASIC test of the VIBE system would be valid for all implementations of this methodology.

Most visualization techniques include the concept of dimensionality, for example, 2D and 3D diagrams, or as in VIBE, expressed by the number of POIs. Adding a new dimension to a visualization system can imply a radical change of the system characteristics. For example, although a 2D scatter plot can be represented directly on a computer screen, we need indirect methods to present a 3D display, for example, using perspective, shadowing, etc. With VIBE, a two-POI display will define a line and a three-POI display a triangle, while the user can define polygon-shaped information spaces with a higher number of POIs. In our case study, we also include the "dimensionality" of scores, which can be Boolean (0 or 1) or vector based (a decimal value).

To control dimensionality in a test situation, we suggest a STEPWISE approach; for example, starting with a simple 2D Boolean case and then moving to higher dimension in successive tests (3D, vector). The 2D case was included as a baseline, reference point for discussing results on progressive tests, even if the visualization method tested would not show its advantage before higher dimensions were introduced. Such a reference point is especially important in the effort to generalize test results. For the testers, a stepwise approach also has the advantage that one may learn as one goes along, that is, the results and experience from the initial tests can be used to streamline the more complicated tests. For our study, the very simple 2D case offered an important foundation for further tests; at the same time it gave confidence in the test methodology. That is, the first test in a stepwise approach should be so simple that one should have a good idea of what the results will be.

This BASSTEP methodology makes it easier to compare the novel visualization method with existing methods, because it is as easy to represent these methods in their BASIC form as the visualization technique. This is, of course, especially the case when initial tests can be performed on paper, but in most cases simple on-line versions can easily be implemented. Again, the idea will be to retain the underlying methodology, not to implement all the bells and whistles of the full systems.

In our case study we have chosen to include text-based, word-based, tabular, and graphical methods, as well as the novel visualization technique. Text-based presentations show words in their usual semantic context. This is the usual form for text lists returned by Internet search engines. Word-based displays show frequency of occurrence of query terms in a document listing. Tables are two-dimensional listings in which the values of the elements are numeric. Graphical displays are the set of usual graph types, for example, pie chart, bar chart, histogram, and scatterplot. Visual displays, in contrast, are composed of icons and connecting lines that do not have the normal Cartesian coordinate interpretation. Icon displays have been used in the TileBars system, and have been observed as a component of some search engine results (Aalbersberg, 1995). The final display is based on the VIBE display using its spring method both in a Boolean and a vector variant.

This distinction is based on work done by Lohse, Rueter, Biolsi, and Walker (1990), investigating how visual displays were categorized. The results of hierarchical clustering analysis of the data showed five clusters—icons, maps, diagrams, network charts, and graph and tables, which closely mirror the four categories of visualization examined in this study.

Case Study

Using our BASSTEP methodology we first establish performance baselines for the simplest retrieval presentations before moving to more sophisticated and novel displays. This provides a control with proper foundation for explaining the possible benefits of the techniques used in the prototype systems.

The initial testing can be performed on paper, which has the advantage of not requiring computer literacy, and eliminating extraneous elements from the test. Paper allows running a large number of subjects, often with only the cost of data entry. The primary drawback with paper tests, in addition to the fact that only static parts of a methodology may be tested, is the difficulty in obtaining timing data on individual tasks. By limiting the time for the whole test, or for parts of the tests, timing data may, to some extent, be measured indirectly through its effect on correctness.

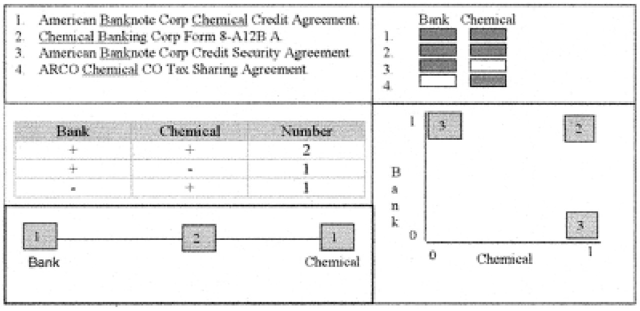
On-line testing is somewhat more costly to prepare than paper tests, but again, we benefit from the defeatured systems, where only the basic functionality has to be implemented. After the test has been prepared, an on-line solution will be cost-independent with respect to the number of subjects. This is especially true of Internet testing, which makes it possible for subjects to perform the test on their own computers. To avoid too great an influence from computer equipment, the test implementation must use a minimum common factor approach, for example, by using a fixed, small window size. A big advantage with on-line tests is that all the test data, including data from a posttest questionnaire, will be collected automatically.

To apply this approach to IR visualization we began with the simplest case of two-term Boolean queries, then studied three-term Boolean queries, and finally investigated more complex vector representations for two and three query terms using on-line testing.

Two-Term Boolean Test (Paper)

The 218 subjects for this study were members of undergraduate courses at University of Pittsburgh or Molde College, Norway. The test was administered as a paper-and-pencil exercise during a normal class meeting. Subjects were given a packet, in English or Norwegian, respectively, containing instructions for completing the experiment, a randomly ordered set of five presentation types, and a posttest questionnaire. The instructions were read aloud to each class before the booklets were opened. Subjects were instructed to refrain from changing answers on a page after they had flipped to the next page. This constraint was applied to more easily detect learning effects over the course of the repeated presentation of questions. No restriction was placed on the amount of time for completing the test, but most subjects handed in their booklet in 10–15 minutes.

Approximately half of the subjects received additional explanation of the various interfaces. The information provided was limited to a preview of each type using dummy data, for example, X and Y rather than actual terms and A and B rather that numeric values.



1. American Banknote Corp Chemical Credit Agreement.
2. Chemical Banking Corp Form 8-A12B A.
3. American Banknote Corp Credit Security Agreement.
4. ARCO Chemical CO Tax Sharing Agreement.

	Bank	Chemical
1.	▨	▨
2.	▨	▨
3	▨	□
4	□	▨

Bank	Chemical	Number
+	+	2
+	-	1
-	+	1

FIG. 3. Sample of presentations types, text list (upper left), icon list (upper right), table (middle left), graph (lower right), and spring/Vibe display (lower left).

The five kinds of presentations were: text, icon list, table, graph, and Vibe display. Figure 3 shows an example of each display condition. Text is ordered so that the items at the head of the list contain both terms, then items containing term X but not Y, and the tail of the list contains Y but not X. The icon list is presented in the same order as text; dark shading indicates the presence of the term and white indicates a term's absence. The table is constructed so that counts of documents containing the various combinations of terms are presented. The graph display plots term X along the X-axis and Y along the Y-axis. The spring display, also called a VIBE display, is based on a model in which documents are placed in a display according to the amount of attraction that the document has for the terms placed at the ends of the line segment. In this two-term instance, documents that are about term "bank" will be counted up at the end of the line labeled "bank." Documents that are about both terms will be counted at the middle of the segment.

For each type of presentation the subject was required to answer two questions: (1) Circle the item(s) that contain terms X and Y; (2) how many items contain the term X?

After all five interfaces had been seen and used by the subject, he was asked to rank the interfaces with respect to: (a) support for answering type A questions; (b) support for answering type B questions; (c) overall preference for general use.

The primary measures of the study are performance and preference. Performance is measured as number of correct answers to the questions related to each display type. In general, preference results concentrate to the subjects' top choice for each ranking category.

To determine if any of the factors probed in the posttest questionnaire might have confounding effects on the study design, we analyzed the data for covariate effects. Overall performance, as measured as total correct answers, or display performance, as measured as the number correct answers per presentation type, was not affected by gender, age, prior computer experience, or current year in academic program. Initial analysis of performance showed a significant effect for country (United States vs. Norway); Norwegian students scored higher on all displays except for the "table" for which performance was equivalent in both groups. Subsequent factoring in of native language resulted in a disappearance of any difference by country in which the study was done. The explanation is that the relatively high

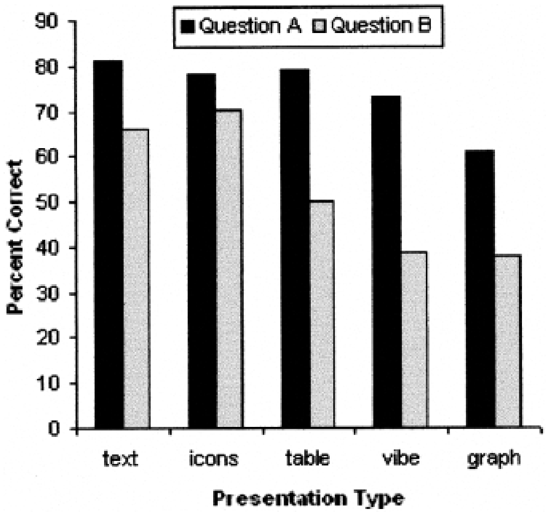


FIG. 4. Overall performance.

proportion of international students in the Pittsburgh sample performed significantly more poorly than the native English speakers. The Norwegian sample did not have any non-native Norwegian speakers.

The tasks that the subjects were asked to perform were chosen to represent two of the Boolean combinations that are possible with a two-term query. Question A corresponds to the logical AND-ing of the terms. Question B is simply the existence of a single term. In all instances Question A was answered correctly more often than Question B for each presentation. The overall performance by Question type is shown Figure 4. The groups of subjects that received an abstract overview of the study performed significantly better than subjects who received only logistical instructions. This was true regardless of question type.

The order of the presentations was randomized to control for order effects. Our results show that a significant amount of learning occurred during the trials. Figure 5 shows that the interfaces that were poorest with respect to performance, i.e., spring model and graph, became more useful if they were presented later in the sequence.

Subjects were asked to rank the five presentations. Overall the icons (by 33%) and the spring model (29%) were considered the best displays, text the worst (by 47%). Sixty percent of the users preferred the visual methods, i.e., icon list and spring displays. It is interesting to note that although performance was superior with the "text" interface, users disliked it.

Three-Term Boolean Tests (Paper and On-line)

A similar setup was used for this test situation. Of the 223 subjects, a paper-and-pencil version of the test was administered to 32 subjects and 191 subjects performed the experiment using the Web. Text, table, icon, and Vibe displays were used in this test.

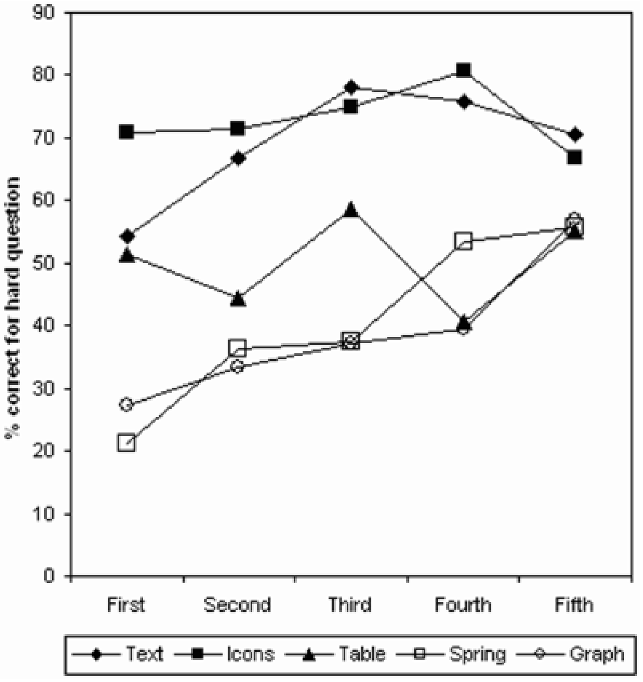


FIG. 5. Percent correctness.

The user's task was to answer questions related to the displays. The same questions were used for each display. The basic form of the questions conformed to a set of Boolean operations using the **and** and **or** connectives. For instance, the question "How many documents have **all** the three terms in them?," is equivalent to A **and** B **and** C.

Performance was assessed as number of correct answers. Computer-mediated sessions were also assessed based on time-to-completion for each display. In addition, subjects were asked to rank the displays with respect to their preference for using them. A posttest questionnaire captured information about the subjects (age, gender, year in program, whether the experiment was performed in their native language), their computer and Internet experience, and, for the computer-mediated group, some specifics about their equipment configuration (modem speed, CPU speed, and monitor size).

The results show that there was no significant difference between computer-mediated administration and paper and pencil. Timing data (Table 3) for the subjects who performed the experiment in the computer-mediated mode showed highly significant differences among the groups when analyzed with a repeated-measures ANOVA.

TABLE 3. Effect of display type on performance (mean + SE).

Display type	Time to answer set of four questions (seconds)	Number correct
Text list	186 ± 9	3.56 ± 0.06
Icon list	175 ± 7	3.64 ± 0.04
Table	147 ± 7*	3.63 ± 0.05
Spring	145 ± 7*	3.35 ± 0.06*

$p < 0.05$ compared to values without an asterisk.

TABLE 4. Preference ratings of various displays.

	Best	Second	Third	Worst
Text	18	17	53	95
Icon list	79	62	31	11
Table	22	57	75	29
Spring	64	47	24	48

Preference results are shown in Table 4. It is clear that the text display was not acceptable to the subjects, while the icon list and spring display were considered very useful.

The second variable that was tested in this study was question difficulty. Table 5 shows the average number of correct responses for each question independent of which display was used in generating an answer. There is a highly significant difference between the levels of difficulty, which is related to the number of **and**s and **or**s that were required. As in the previous two-term study, questions requiring the use of **or** were more difficult. In this case, Question #2 was phrased so that it required the subject to use an **or**.

Vector Studies (On-line)

Vector studies were performed with displays as shown in Figure 6. For the text display all words except keywords were filtered, while the table gave the number of occurrences for each keyword. One hundred ninety-five subjects were randomized to receive either the two-term or three-term experimental study. Both vector studies were performed on-line.

To determine whether the two performance measures employed in this study—time to completion, and number of correct answers—were correlated, the data for overall test performance on both scales over all displays was analyzed visually and statistically. Figures 7 and 8 show the results for the two-term and three-term studies, respectively. Open squares indicate outliers. Diamonds show data for the remaining subjects. The trendline shows association between measures for diamonds. This comparison shows that the primary measures used in this study are not correlated. In other words, performance measured by time to complete a task is not predictive of the score that the subject is likely to achieve. Subjects who completed the total battery of tasks in a relatively short amount of time were no more likely to achieve a high score than subjects who took longer. Similarly, subjects who scored particularly well or particularly poorly were not associated with skewed performance times. The Pearson Correlation Coefficient was .038 and .177 for two-term and three-term data, respectively; neither value was statistically significant.

From the distribution of values, it appears that time exhibits a wider range of values while correctness is more constrained. An inference could be made that time is a more sensitive measure. It might also be suggested that the type of test that was administered was quite easy, and that subjects performed too well to allow correctness to be discriminating.

Subjects were presented with a short description of an upcoming display type. The material consisted of an explanation of the key elements in the display and an example of how it could be interpreted. When the subject was finished using this information, he submitted a request for the first display of this type. The time elapsed was captured and labeled as “instruction time.” Table 6 shows a statistical summary of the data. On average, the instructional material was viewed for less than a minute. The amount of time spent learning about a display was similar for the word, icon, and table display and the triangular “spring” display used in the three-term study. In the two-term study, both the graph and the linear “spring” required significantly longer times. These longer times seem to indicate a degree of novelty of the displays. The fact that the thee-term “spring” was not accompanied by a longer instructional period would not be expected. It might be conjectured that the “triangle” was less confusing than its “linear” counterpart, but no data was gathered that could support or refute this idea.

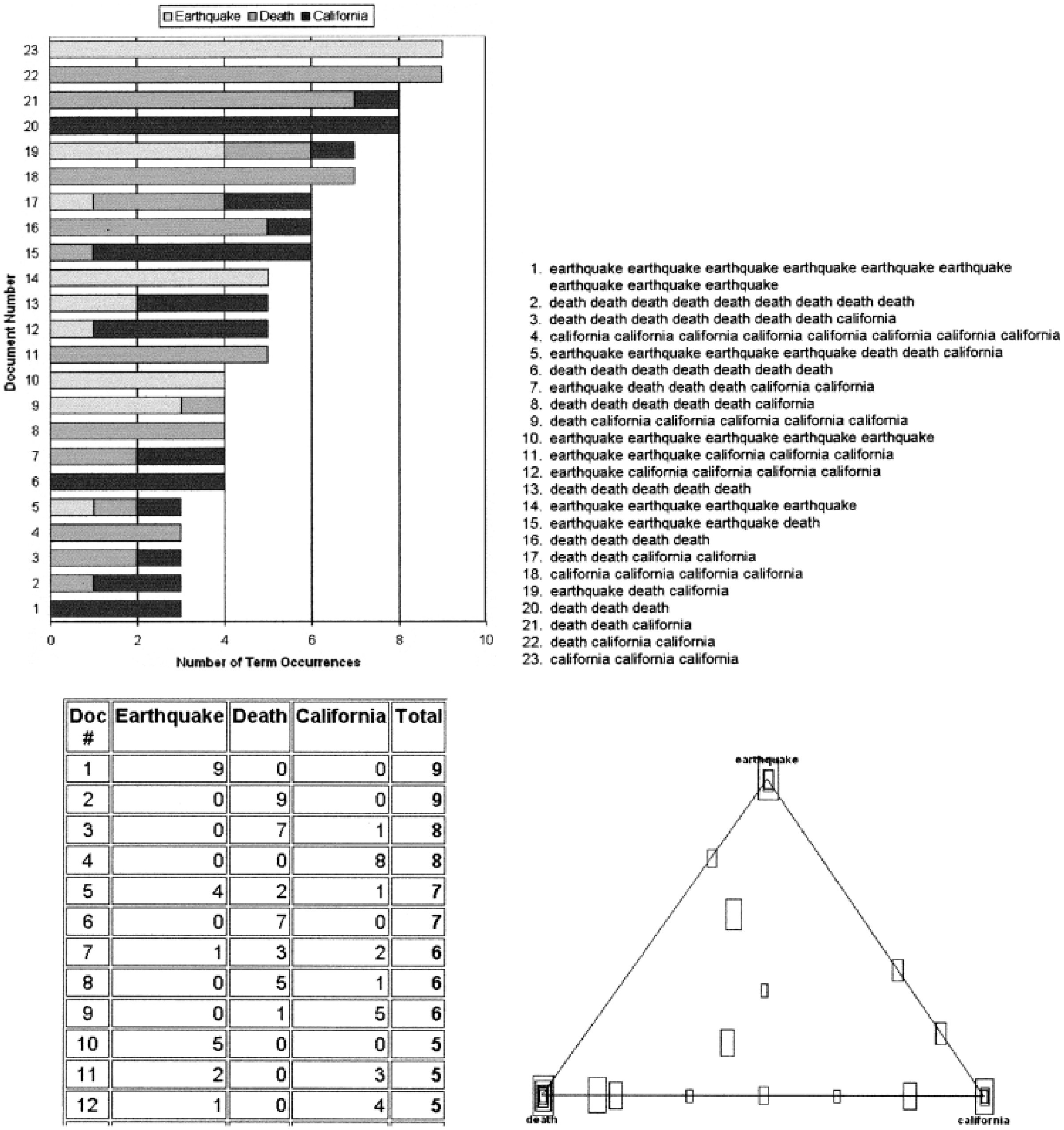
The results of the analysis of time to completion with respect to display type are shown graphically in Figure 9. There are several important observations that can be made upon inspecting the data. There are significant differences among both the two- and three-term displays with respect to performance times. Analysis of variance showed a p value $< .001$ for this comparison. Data on correctness provided similar information; time to completion, however, appeared to be more sensitive. The variation in timing data, assessed as the standard error of the mean, was larger than the relative standard error for correctness data.

It was shown that each display type was associated with poorer performance, as measured by time to completion, when it was presented first in the series. This effect was not shown for the number of correct answers. The key observations regarding the time effect are: (1) there is a steep drop in time required between the first and second display regardless of which displays were seen in these slots; and (2) the “spring” display is handled extremely rapidly in the three-term condition; the “spring” display is the only display that is not influenced by the increased complexity of the three-term condition when compared with the paired two-term display. Perhaps this shows that the idea behind the more complex spring display is more clearly understood in the more complex situation.

Subjects ranked the displays after using all of them. Analysis showed that there was no relationship of these

TABLE 5. Performance as a function of question type.

Question no.	Composition	Correct answers (mean ± SEM)
1	3 ands	3.8 ± 0.3
2	3 ands + 1 or	3.1 ± 0.2
3	4 ands	3.5 ± 0.3
4	3 ands	3.6 ± 0.3



1. earthquake earthquake earthquake earthquake earthquake earthquake earthquake earthquake earthquake
2. death death death death death death death death death
3. death death death death death death death california
4. california california california california california california california california
5. earthquake earthquake earthquake earthquake death death california
6. death death death death death death death
7. earthquake death death death california california
8. death death death death death california
9. death california california california california california
10. earthquake earthquake earthquake earthquake earthquake
11. earthquake earthquake california california california
12. earthquake california california california california
13. death death death death death
14. earthquake earthquake earthquake earthquake
15. earthquake earthquake earthquake death
16. death death death death
17. death death california california
18. california california california california
19. earthquake death california
20. death death death
21. death death california
22. death california california
23. california california california

Doc #	Earthquake	Death	California	Total
1	9	0	0	9
2	0	9	0	9
3	0	7	1	8
4	0	0	8	8
5	4	2	1	7
6	0	7	0	7
7	1	3	2	6
8	0	5	1	6
9	0	1	5	6
10	5	0	0	5
11	2	0	3	5
12	1	0	4	5

FIG. 6. Examples of icon, test, table, and spring displays used in the three-term vector study.

preference rankings and subject performance, when measured by time to completion. There was, however, a correlation between rankings and correctness for both the two-term and three-term groups. In each case, subjects who received high scores when using the "spring" display preferred it. In the two-term study, the same observation was made for Graph.

In addition to ranking the display, the subjects were given the opportunity to rate the displays as "Easy," "Hard," "Fun," and/or "Annoying." Every subject voted in at least one category, and many people selected more than one display as exhibiting a certain characteristic. The percentages are shown in Table 7 for the two-term study, Table 8 for the three-term.

These data confirm the results of the rankings. As the difficulty of the scenario increased, i.e., two-term to three-term condition, the Word display became significantly more difficult to use (50% of two-term subjects vs. 78% of three-term subjects), while the "spring" display became more useful (i.e., significantly easier, less hard, more fun, and less annoying). The "spring" display was perceived in the harder environment to be easier and more fun to use.

Comparison of Two- and Three-Term Studies, Boolean, and Vector, Paper and On-line

The primary hypothesis that was being tested in this experiment was that the enhanced difficulty of the setting

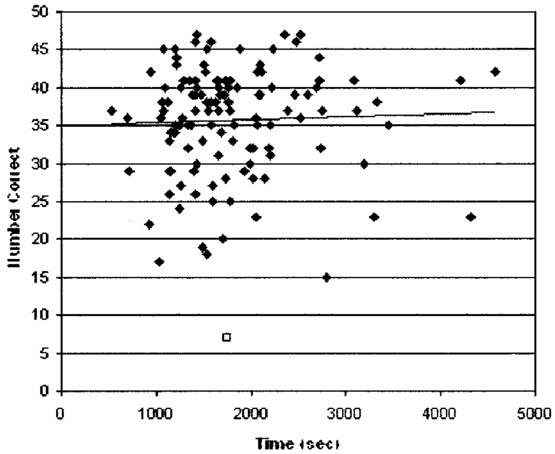


FIG. 7. Relationship of time to completion and correctness for two-term study.

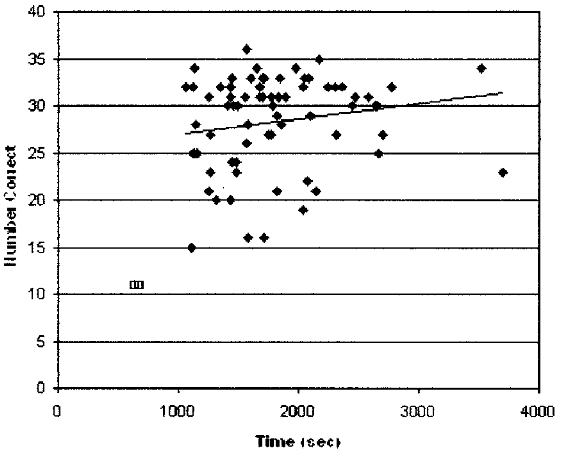


FIG. 8. Relationship of time to completion and correctness for three-term study.

TABLE 6. Time spent on instruction page (seconds; mean ± SEM).

	2-Term (n = 120)	3-Term (n = 72)
Word	35.35 ± 2.91	39.30 ± 4.39
Icon	35.17 ± 2.46	32.93 ± 3.05
Table	35.60 ± 2.92	32.49 ± 3.04
Graph	54.38 ± 3.44*	NA
Spring	51.76 ± 4.55*	35.14 ± 3.14

* $p < 0.05$ when compared with other two-term displays.

(two-term vs. three-term Boolean, Boolean vs. vector) would show a superior performance with visual displays. This immunity to performance decay would be accompanied by an increased preference of subjects for the visual displays.

A cross-study comparison was performed for the two- vs. the three-term Boolean study. To compare the studies, the data were adjusted by removing references to the Graph presentation in the two-term study. The Kruskal-Wallis test was applied to the resultant data, and it showed that the rankings for best and for worst display were significantly different (Table 9). The inference than can be drawn from this data is that the "spring" display was preferred more often in the more difficult three-term study than in the easier two-term condition. This was confirmed by the vector studies. As the difficulty of the scenario increased, i.e., two-term to three-term condition, the Word display became significantly more difficult to use, while the "spring" display became more useful. The "spring" display was perceived in the harder environment to be easier and more fun to use. However, the more familiar icon and table displays were considered the easiest to use.

Discussion

As illustrated by this study, the BASSTEP approach has several advantages for testing of novel techniques:

(1) Tests, especially on paper, are simple to prepare.
(2) Results from each test are simple to analyze.
(3) Results may be used to plan the next step.
(4) Conclusions may be drawn based on a comparison of results from different steps.
(5) The simplicity of the tests, paper or on-line, make it practical to use a large number of subjects, thus decreasing the influence from outside factors and increasing confidence levels.
(6) Comparison to other basic methodologies, not to full systems. That is, the comparative data will be simpler to interpret and of more general value, than if the test had included representatives of full systems.
(7) Results will be valid for all systems that utilize the visualization methodology tested.

We found few differences in results between paper and on-line tests, showing that the choice of medium for these types of tests may safely be made based on pragmatic considerations alone. Paper tests are simpler to prepare than the on-line tests, but on-line testing can be fully automated, allowing large numbers of subjects to be run conveniently. On-line testing has the additional advantage of allowing collection of timing data. This allowed us to draw more and better conclusions than for the paper data, where only cor-

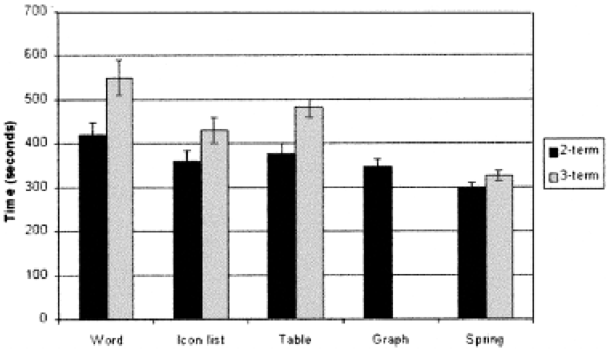


FIG. 9. Time to completion.

TABLE 7. Two-term ($n = 120$) study, percentage of subjects categorizing display according to various criteria.

	Easy	Hard	Fun	Annoying
Word	10	50	8	74
Icon	51	9	34	6
Table	39	7	7	21
Graph	48	13	30	17
Spring	10	55	15	50

rectness was used as a dependent measure. Although a step-wise approach may be applicable in many testing situations, this principle is most useful for new, nonmature areas. The hundreds of different visualization systems proposed is a clear sign that we are still in "let the hundred flowers bloom" phase of development. With the methodology proposed here, it becomes possible to test the usability of visualization techniques. Through the BASSTEP methodology we can evaluate and compare the visualization principles behind the systems, without considering the bells and whistles of the systems in which they were implemented. The BASSTEP method allows comparisons that are independent of the resources spent on implementing prototypes or systems. This is especially applicable to visual systems, as these often rely on one or a small set of basic visualization strategies to add value to a complex information processing system.

From observable usability, and natural selection, we expect that certain of these visualization methods will be accepted as the better, and that new systems will emerge that rely on these techniques. When a set of consensus techniques is established, the differences will become more between systems than techniques, and the BASSTEP method will no longer be applicable. This is clearly true in more mature areas such as spreadsheet systems. Today these systems rely on the same basic principles, and the differences between them are found in details of implementation rather than the conceptual design of the system.

Conclusion

This study illustrates a systematic approach to evaluating novel information displays. By isolating basic representational constructs from the system implementation and defining tasks independently of existing systems, it becomes possible to evaluate new interface techniques directly without the delay and effort of full implementation and deployment. This stepwise method (BASSTEP) was applied to both paper and on-line tests. Although the data showed no significant differences between these two media, on-line tests gave the possibility of adding more accurate timing data to the study, as well as automating the data collection process. In some conditions, however, the advantages of on-line testing may be outweighed by the greater difficulty of implementing them.

BASSTEP provided clear results in a case study where five different display types used in existing information retrieval visualization systems were evaluated. Of text, tables, icon, graphs, and a novel visualization technique, the spring/Vibe technique, the graphical methods (icon, graphs, "spring") were preferred when the complexity of the task increased. The novel method proved to give best performance in the most complex situation, a three-term task with vector data.

TABLE 8. Three-term ($n = 72$) study, percentage of subjects categorizing display according to various criteria.

	Easy	Hard	Fun	Annoying
Word	3	78**	4	89*
Icon	56	3	17*	4
Table	33	7	7	17
Spring	29**	24**	47**	21**

* $p < 0.05$; ** $p < 0.01$.

TABLE 9. Results of Kruskal-Wallis analysis of ranking data with respect to study type.

	Best	Second	Third	Worst
Chi-square	6.308	1.389	2.187	26.746
Significance	0.012	0.239	0.139	0.000

Acknowledgments

This work was supported by a grant from the National Science Foundation (#9529072).

References

Aalbersberg, I.J. (1995). Personal communication in Nuchprayoon (1996).

Au, P., Carey, M., Sewraz, S., Guo, Y., & Ruger, S.M. (2000). New paradigms in information visualization. Proceedings of SIGIR'2000, Athens, Greece (pp. 307–309).

Chalmers, M., & Chitson, P. (1992). Bead: Explorations in information visualization. Proceedings of SIGIR'92, Denmark (pp. 330–337).

Crouch, D., & Korfhage, RR. (1990). The use of visual representations in information retrieval applications. In T. Ichikawa, E. Jungert, & R.R. Korfhage (Eds.), Visual languages and applications (pp. 305–326). New York: Plenum Press.

Fox, K., Frieder, O., Knepper, M.M., & Snowberg, E.J. (1999). SENTINEL: A multiple engine information retrieval and visualization system. Journal of the American Society for Information Science, 50(7), 616–625.

Hearst, M.A. (1994). Using categories to provide context for full-text retrieval results. Proceedings of the RIAO '94, New York.

Hearst, M.A. (1995). TileBars: Visualization of term distribution information in full text information access. CHI '95 Proceedings (pp. 213–220).

Hearst, M.A., & Karadi, C. (1997). Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. Proceedings of SIGIR'97, Philadelphia, PA (pp. 246–254).

Kakimoto, T., & Kambayashi, Y. (1999). Browsing functions in three-dimension space for digital libraries. International Journal on Digital Libraries, 2, 68–78.

Kim, H., & Korfhage, R.R. (1994). BIRD: Browsing interface for the retrieval of documents. Proceedings of the 1994 IEEE Symposium on Visual Languages, St. Louis, MO (pp. 176–177).

Korfhage, R.R. (1997). Information Storage and Retrieval. New York: Wiley.

Korfhage, R.R, & Olsen, K.A. (1991). Information display: Control of visual representations. IEEE Workshop on Visual Languages, Kobe, Japan (pp. 56–61).

Koshman, S. (1996). Usability testing of a prototype visualization-based information retrieval system. Dissertation, University of Pittsburgh.

Lin, X. (1996). Graphical table of contents. Digital Libraries, DL'96, Bethesda, MD (pp. 45–53).

Lohse, G., Rueter, H., Biolsi, K., & Walker, N. (1990). Classifying visual knowledge representations: A foundation for visualization research. Visualization '90: Proceedings of the First Conference on Visualization (pp. 131–138).

Mackinlay, J.D. (1986). Automating the design of graphical presentations of relational information. ACM Transactions on Graphics, 5(2), 111–141.

Morse, E., & Lewis, M. (1997). Why information visualizations sometimes fail. Proceedings of IEEE International Conference on Systems Man and Cybernetics, Orlando, FL, October 12–15, 1997.

Mothe, J., & Dkaki, T. (1998). Interactive multidimensional document visualization. Proceedings of SIGIR'98, Melbourne, Australia (pp. 363–364).

Newby, G.B. (1992). An investigation of the role of navigation for information retrieval. Proceedings of ASIS '92 (pp. 20–25).

Nuchprayoon, A. (1996). GUIDO: A usability study of its basic retrieval operations. Doctoral Dissertation. School of Information Sciences, University of Pittsburgh.

Olsen, K.A., Williams, J.G., Sochats, K.M., & Hirtle, S.C. (1992). Ideation through visualization: The VIBE system. Multimedia Review, 3(3), 48–59.

Sebrechts, M.M., Vasilakis, J., Miller, M.S., Cugini, J.V., & Laskowski, S.J. (1999). Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. Proceedings of SIGIR'99, Berkeley, CA (pp. 3–10).

Shaw, C.D., Kukla, J.M., Soboroff, I., Ebert, D.S., Nicholas, C.K., Zwa, A., Miller, E.L., & Roberts, D.A. (1999). Interactive volumetric information visualization for document corpus management. International Journal on Digital Libraries, 2, 144–156.

Spoerri, A. (1993). InfoCrystal: A visual tool for information retrieval. Proceedings Visualization '93, San Jose, CA (pp. 150–157).

Swan, R.C., & Allan, J. (1998). Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. Proceedings of SIGIR'98, Melbourne, Australia (pp. 171–181).

Veerasamy, A., & Belkin, N. (1996). Evaluation of a tool for visualization of information retrieval results. Proceedings of SIGIR'96, Zurich, Switzerland (pp. 85–92).

Veerasamy, A., & Heikes, R. (1997). Effectiveness of a graphical display of retrieval results. Proceedings of SIGIR'97, Philadelphia, PA (pp. 236–245).

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

Zhang, J., & Korfhage, R.R. (1999). A distance and angle similarity measure method. Journal of the American Society for Information Science, 50(9), 772–778.