

Cluster-based and Association-based Visualization Systems as Information Exploration Tools

Min Song

School of Library and Information Science
Indiana University, Bloomington, Indiana

Abstract

The main purpose of this study is to investigate whether two of the algorithms, variance of Ward's hierarchical clustering algorithm and a Kohonen neural network algorithm, can help improve information exploration of unknown data collections. The initial results of the study indicate that both BiblioMapper-based and Kohonen SOM-based algorithms can successfully categorize heterogeneous data collections into manageable sub-spaces that users can successfully navigate to locate a document of interest. Both BiblioMapper and Kohonen SOM worked best with browsing tasks that were very broad, in which subjects skipped around between categories. Subjects who preferred keyword search and those who wanted to use the more familiar mental models (alphabetical organization) for browsing found that the map did not work well.

Introduction

Visualization is a cognitive process performed by humans in forming a mental image of a domain space. According to Williams et al. (1995), particularly in information science, visualization is "the visual representation of a domain space using graphics, images, animated sequences and sound augmentation to present the data, structure and dynamic behavior of large, complex datasets that represent systems, events, processes, objects and concepts (p. 163)". Current visualization approaches demonstrate effective methods for visualizing structured and/or hierarchical information such as organization charts, directories, entity-attribute relationships, etc. Free text visualizations have remained relatively unexamined.

At the same time, "open source" digital information--the kind available freely or through subscription over the Internet--is increasing exponentially. Whether the purpose be market analysis, environmental assessment, law enforcement or medical analysis, the task is to peruse large amounts of text to detect and recognize informational "patterns" and pattern irregularities across the various sources. But modern information technologies have made so much text available that it overwhelms the traditional reading methods of inspection, sifting and synthesis.

True text visualizations that would overcome these time and attentional constraints must represent textual content and meaning to analysts without them having to read it in the manner that text normally requires. These visualizations would instead result from a content abstraction and spatialization of the original text document that transforms it into a new visual representation

that communicates by image instead of prose. Then the image could be understood in much the way that we explore our worldly visual constructions.

I have developed a visualization system of document space named BiblioMapper and have demonstrated that BiblioMapper can successfully be used as a stand-alone information exploration tool for CISI data collections (Song, 1998). The purpose of the current study is to investigate the influence of document classification methods on information visualization and evaluate the visualization systems. In this study, a clustering algorithm adapted in BiblioMapper and a neural network algorithm, Kohonen Self-Organizing Map (SOM), are used for visualization of a document space.

Statement of the problem

The primary requirement of a text processing engine for information visualization are: 1) the identification and extraction of essential descriptors or text features; 2) the efficient and flexible representation of documents in terms of these text features; and 3) subsequent support for information retrieval and visualization. A number of textual analysis techniques have been introduced to identify descriptors and develop an unambiguous internal representation for a document. Automatic indexing in information science and natural language processing in artificial intelligence are two sets of techniques frequently used for textual analysis. BiblioMapper and SOM were constructed based on term-frequency-based automatic indexing technique.

The most important but relatively unexplored part of information visualization is how effectively the documents are represented in terms of these text features. Most of the visualization systems of document space depend on neural network algorithms (Lin et al., 1991; Kohonen et al., 1996; Belew, 1989). Kohonen SOM, chosen for the study, has been rigorously applied to textual document classification (Orwig et al., 1997; Lin et al., 1991). In the comparison with the application of neural network algorithms to information visualization, classical clustering algorithms have not been proactively used for visualization. In previous study, the application of hierarchical clustering algorithms to information visualization was theoretically justified (Song and Gillespie, 1997).

Document representations by these two classification methods are unique: one is cluster-based and the other is association-based. In the current study, the research focus is on investigating usability of these two visualization systems of a document space, stemming from two different classification methods.

Document classification in BiblioMapper and Kohonen map

BiblioMapper

The clustering algorithm used in BiblioMapper is based on a minimum distance pairs technique similar to Ward's minimum variance clustering algorithm (Ward, 1963). Minimum distance pairs are established by keeping track of both the nearest neighbor of each data point and the distance.

The nearest neighbors are initially computed in $O(N^2)$ time. It is assumed that the similarity between the chosen documents is the inner product of two vectors using appropriately weighted vectors. In this case, the similarity between a cluster centroid and any other document is equal to the mean similarity between the document and all the other documents in the cluster. Since the centroid of the cluster is the mean of all the document vectors, the centroid can be used to compute the similarities between the clusters while requiring only $O(n)$ space. When a new cluster is formed, one of the child nodes is discarded, while the other is replaced by the new node to represent the new cluster. Each of the other points is checked to see whether the nearest neighbor is re-constructed by the new cluster.

The visualization technique used in the BiblioMapper system is based on the multidimensional scaling of the similarities between the inner-class vectors and dissimilarities between the outer-class vectors generated from the minimum distance pairs algorithm to plot one point in Euclidean space for each document. These points form a semantic scatterplot by clustering similar documents around a centroid. The degree of similarity, associated with topical relevance, is determined by the distance of vectors from the centroid. The centroid represents the subjectivity of each class. Since semantic relationships are represented by this distance and there might be some documents with the same distance from a centroid, plotting the points with the same distance poses a problem. In order to overcome this problem, the angle of two vector points are computed by cosine formula (Salton and McGill, 1983).

$$\text{theta}(\theta) = \frac{\text{sum}}{\sqrt{\text{sum1} * \text{sum2}}}$$

BiblioMapper represents each document as a point in two-dimensional space. The distance between two representative points--one a document and the other is centroid--is roughly proportional to the corresponding vectors. This makes it possible for the documents of a class to be mapped out close to the centroid in two-dimensional space, much like books on the shelves of a library arranged by some classification scheme. Kohonen's map visualizes the document space in a manner similar to BiblioMapper by dividing the available space into semantic regions (Kohonen, 1996).

Kohonen's SOM map

An approach to Kohonen's map algorithm is an unsupervised learning method well-known in artificial neural networks (Kohonen, 1989). The algorithm takes a set of input objects, each represented by an N-dimensional vector, and maps them onto nodes of a two-dimensional grid. Kohonen's algorithm contains two layers of nodes--an input layer and a mapping (output) layer in the shape of a two-dimensional grid. The input layer acts as a distribution layer. The number of nodes in the input layer is equal to the number of features or attributes associated with the input. Each node of the mapping layer also has the same number of features as there are input nodes. Thus, the input layer and each node of the mapping layer can be represented as a vector which contains the number of features of the input. The network is fully connected in that every mapping node is connected to every input node. The mapping nodes are initialized with random numbers. Each actual input gets compared with each node on the mapping grid. The winning mapping node is designed as that with the smallest Euclidean distance between the mapping node vector and the input vector. The input thus maps to a given mapping node. The value of the

mapping node vector is then adjusted to reduce the Euclidean distance. In addition, all of the neighboring nodes of the winning node vector are adjusted proportionally. In this way, the multi-dimensional (in terms of features) input nodes get mapped to a two-dimensional output grid. After all of the input is processed (usually after hundreds or thousands of repeated presentations), the result should be a spatial organization of the input data organized into clusters of similar (neighboring) regions.

Several recent studies adopted the SOM approach to textual analysis and classification. Ritter and Kohonen (1989) apply the Kohonen SOM to textual analysis in an attempt to detect the logical similarity between words from the statistics of their context. In support of using Kohonen for textual document classification, Lin et al. (1991) was the first to adopt the Kohonen SOM for information retrieval. In his prototype, self-organizing clusters of important concepts in a small database of several hundreds of documents were generated. A scaleable multi-layered, graphical SOM approach to Internet categorization was developed by Chen et al. (1996) and the resulting prototype was tested for usability in a second experiment (Chen et al., 1998).

The system

The study uses a randomly selected group of subjects and a representative sample of documents consisting of journal abstracts, e-mail communications and other internet resources drawn from the domain of information science (IS).

The database being tested with BiblioMapper and Kohonen SOM contains Internet resources related to computer science and information science (800 document surrogates). The BiblioMapper was designed to function in the X window environment. The system was built and run on a Silicon Graphics Indigo Extreme. Tcl/tk was used to build up interfaces of BiblioMapper. The presented version of Kohonen SOM was developed by a research group of the artificial intelligence laboratory at the University of Arizona.

Figure 1 presents a typical view of the BiblioMapper interface. There are 19 clusters, with each cluster divided by a rectangular area. For each cluster, the user is presented with the topical words that represent the text contents of the documents in a cluster. A list of topical words of a cluster is shown in the pop-up menu window by clicking on a cluster number. If the user wants to look into one of the clusters, the selected cluster is opened by pointing the mouse cursor onto the cluster. Figure 2 shows the window of the selected cluster. For the user who wants to see the title of a document first, the title window is opened up by double-clicking on an icon (Figure 3). When the user wants to see the document surrogate, by double clicking with the second button of the mouse on the icon pulls it up on the screen (Figure 4). Users judge relevance of documents according to a scale whose range is from 0 to 10 that is located at the bottom of the document window. Relevance judgments are saved in a log file. Figure 5 presents a typical view of the Kohonen SOM interface. The user is allowed to search the database either with an alphabetical subject list or with a map display. If the user clicks the subject of interest, a list of documents is brought up to allow the user to locate the document of interest by browsing the list.

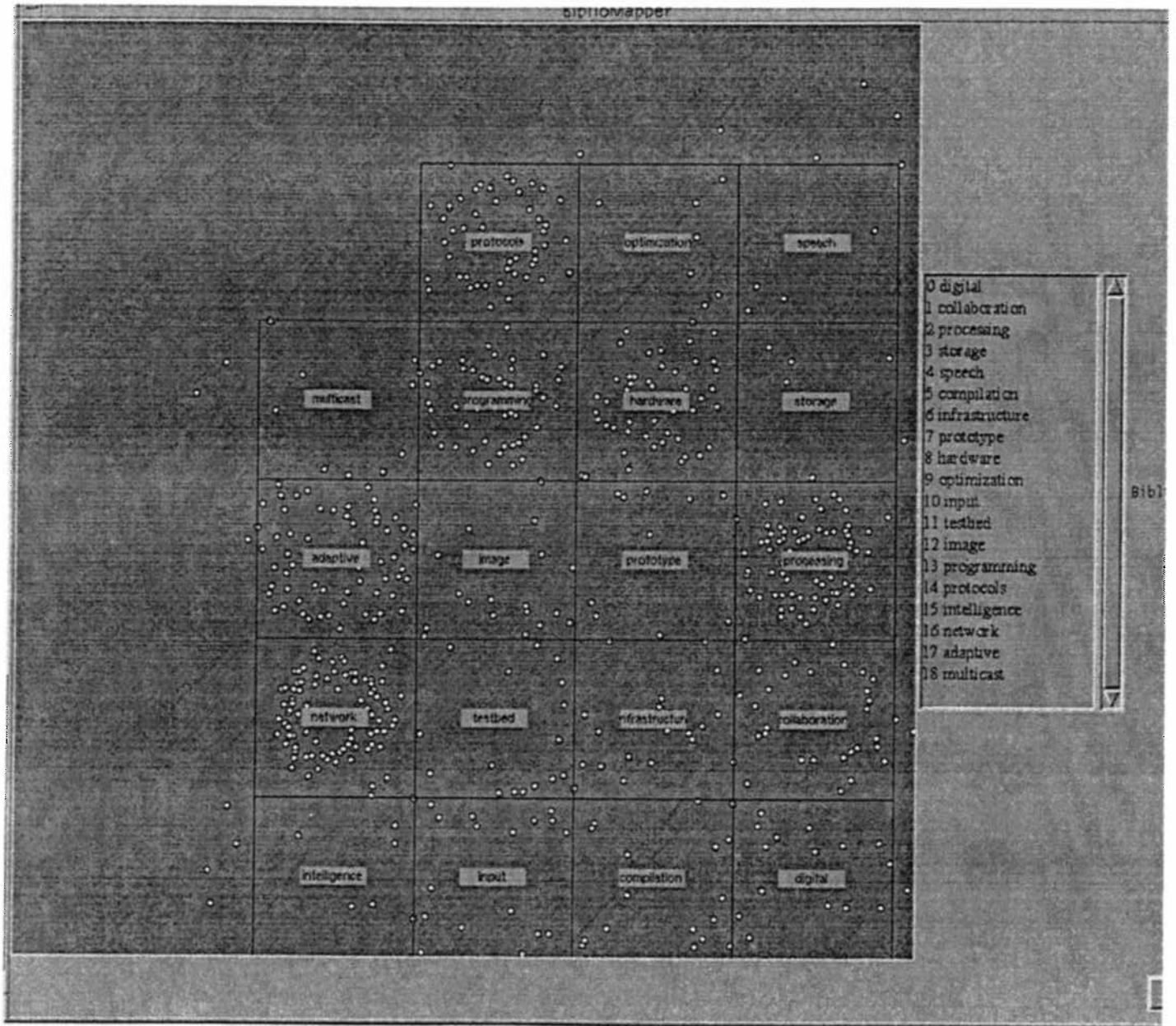


Figure 1. A Snapshot of the Bibliomapper Interface

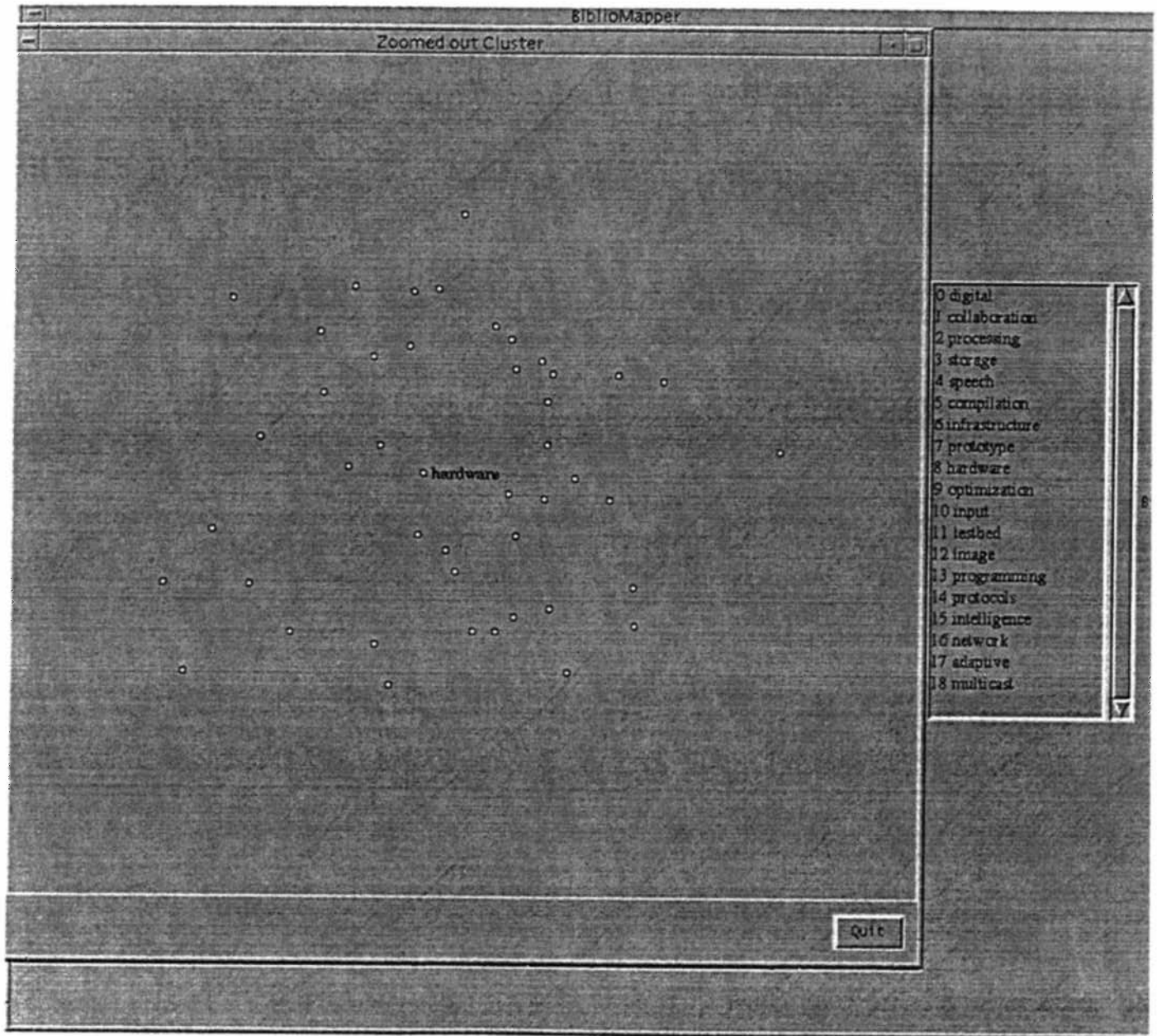


Figure 2. A Snapshot of the Cluster Window

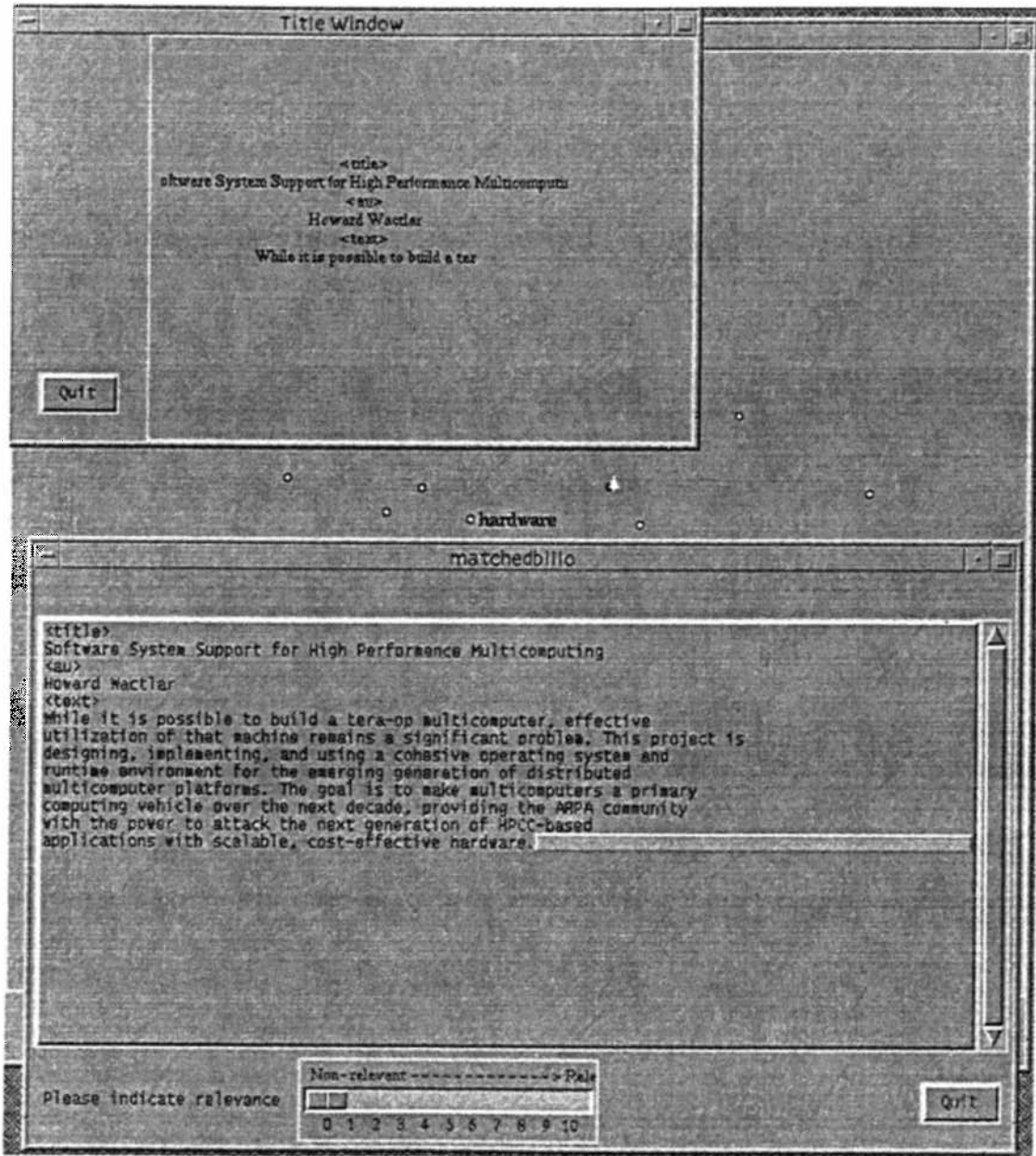


Figure 3. A Snapshot of the Title Window

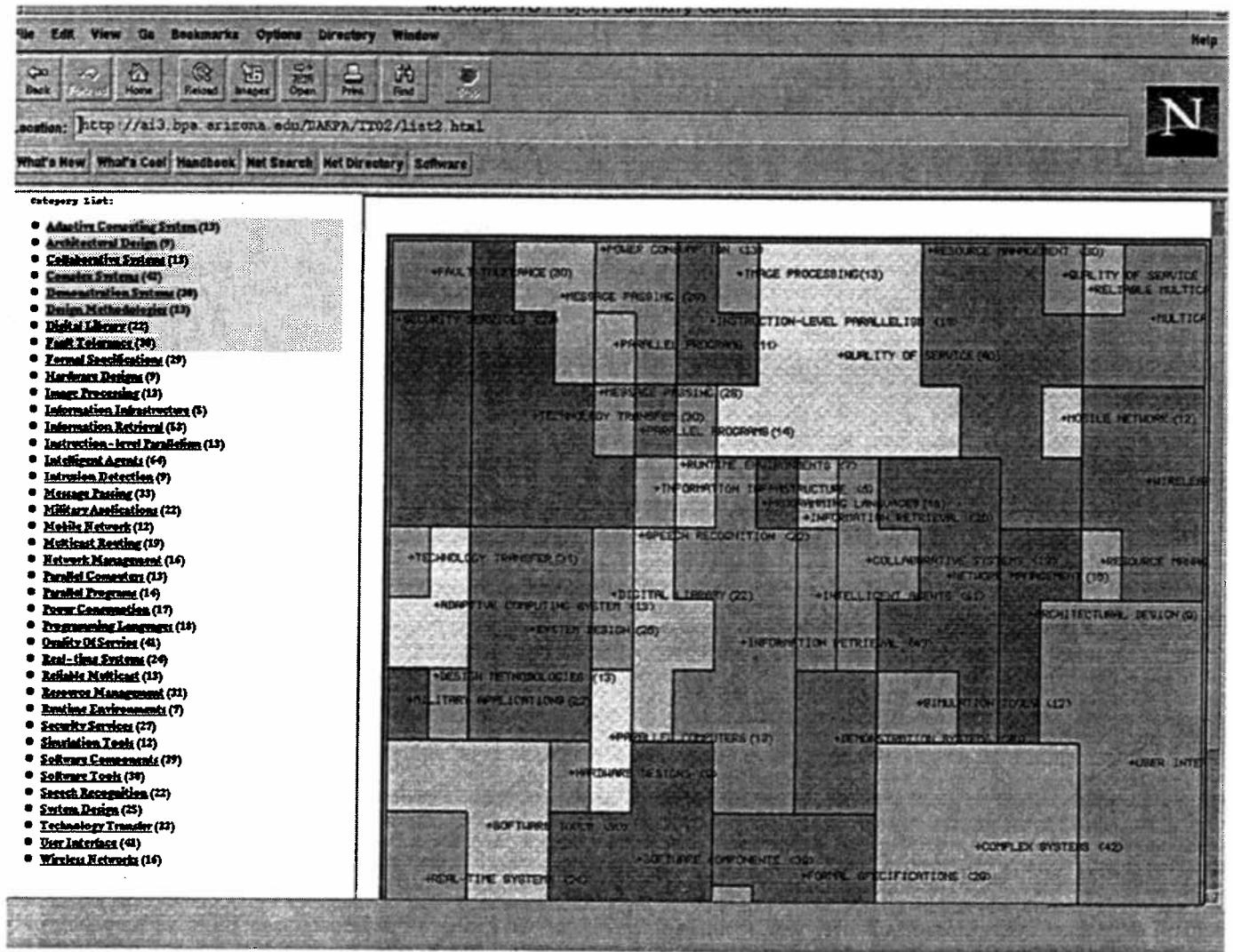


Figure 4. A Snapshot of the Kohonen SOM Interface

The experiment

Evaluation of visualization systems has been a major challenge in the situation where recall and precision have been shown to be inadequate to capture the rich variety of tasks typical of information exploration interactions (Meadow, 1992). A variety of new measures derived from recall and precision have been used (Veerasamy and Heikes, 1997; Golovchinsky, 1997) with partial success. In this study, evaluation of the system is limited to the current research goal: correspondence of the visualization scheme to the users' mental model. The current study concentrates on comparisons of BiblioMapper and SOM, with respect to their effectiveness as information exploration tools. The corrected precision measure, proposed by Janes (1991), is used for measuring search accuracy.

Participants

Ten adults participated in this study. Seven of the participants were undergraduate students at Indiana University majoring in Arts and Humanities programs (three female and four male, age range: 17-24). Three of the participants were graduate students studying Music (one male, age 27 and two females, age 28 and 30). All of the participants were not familiar with online searching. In particular, they reported that they had no experience with visual interfaces for IR systems.

It should be noted here that the user type variable manipulated expertise in searching with IR systems, not domain expertise of library and information science. The domain variable is likely to be worthy of further investigation but is not central to the focus of this system design.

Experimental Design

A two-factor design compared all users on both systems - BiblioMapper and Kohonen SOM. Independent variables were query difficulty (easy, medium, and difficult), and visual interface style. Dependent variables were time taken to complete tasks, accuracy of search performance, and navigation style (operationalized as number of nodes visited in the system) and responses to a post-task interview.

Subjects were required to complete three information tasks (see appendix). These tasks were developed by the author to ensure that tasks did not unduly favor any one medium and that users would need to explore the system fully to complete the tasks.

Materials and Procedure

All subjects were tested individually at a single workstation located in Indiana University's SLIS library. Before the search session, brief instructions and training were given by the author. At this session, the author described the nature of the investigation and introduced the subjects to the systems and tasks for 20 minutes. The subjects were told that they would be interrupted between tasks to allow the author to ask some questions related to the search. Subjects were then given a set of tasks and asked to attempt all tasks in the order presented. The subjects were encouraged to verbalize their thoughts. Their comments and movement through screens were recorded by the author.

RESEARCH FINDINGS

A series of analyses concentrated on comparisons of BiblioMapper and Kohonen SOM with respect to their effectiveness as information browsing tools. Results are presented below for each category of data: completion time, accuracy, navigation, and structured interviews.

Finding Relevant Articles

Table 1 presents a summary of data for the mean time spent per query, the mean number of documents selected for retrieval, and the mean *corrected precision* (Janes, 1991) of the retrieved documents. Corrected precision takes account of differences between users in their selectivity when judging the relevance of documents. The formula for corrected precision, proposed by Janes (1991), is as follows:

$$BP = \frac{\sum_{k=1}^n R_k}{N}$$

P is the precision ratio. In this study, P is calculated by dividing sum of relevance judgments by number of documents viewed. BP is the break point expressed by the user. In this study, BP is calculated by the following formula:

$$CP = 100 \frac{P}{100 - BP}$$

R is relevance judgments whose ranges are between 0 to 10 and N is number of documents with relevance given by the user

For each type of data in table 1, (i.e., each row) I conducted an analysis of variance (ANOVA) for the two groups (BM, SOM) x Query Difficulty (Easy, Medium, Hard).

	BiblioMapper (BM)	Kohonen SOM (SOM)
Time (min.)	23.10	21.51
Corrected Precision	.54	.49
No. Selected Documents	27.26	27.96

Table 1. Pre-query average for BM and SOM groups.

Completion Time

Examining Table 2, it is apparent that completion time of participants who used the BiblioMapper interface to answer queries was equal to those using Kohonen SOM. The BM vs SOM group difference in Table 2 was not significant, $F(1, 59) = 2.039, p > .05$.

Source of variation	Sum of Square	df	Mean of Square	F-ratio
Interface Styles	34.144	1	38.144	2.039
Query Difficulty	873.238	2	436.664	23.344*
Interaction	42.806	2	21.403	1.144
Within Cell	1010.112	54	18.706	
Total	1964.390	59	33.295	

Table 2. Two-way ANOVA analysis of completion time

Time taken for browsing increased with increase in Query difficulty. Query difficulty had a significant effect, $F(2,59) = 23.34$, $p < .05$, with participants spending more time on Difficult ($M=27.69$ min.) and Medium ($M=19.23$ min.) queries than on Easy queries ($M=20.01$ min.). This main effect did not interact with the type of interface used, $(2, 59) = 1.14$, $p > .05$. The subjects who conducted difficult information tasks with the BiblioMapper interface spent 27.6 minutes, and the subjects with the Kohonen SOM interface spent 27.77 minutes..

Query Difficulty	Interface			
	BiblioMapper		Kohonen SOM	
	Mean	SD	Mean	SD
Easy	20.56	2.29	19.48	7.43
Medium	21.17	3.44	17.29	4.01
Difficult	27.60	3.70	27.77	3.14

Table 3. Mean times and standard deviation per task (in minutes).

Accuracy

Summary data is presented in Table 4, where maximum mean value in each category is 1. The difference between the two interfaces is not significant. BiblioMapper performed slightly better than Kohonen SOM. However, a significant effect was found for query difficulty ($F[2, 59] = 30.267$, $p < .05$). The search performances of the users with both interfaces were poor in terms of the corrected precision measure (Table 5). This may be due to the fact that none of the users had domain knowledge in computer and information science. In addition, search tasks assigned to the users resulted in user's poor searching performance (it was revealed by the post-interviews that the subject would be more interested in searching if the domain had been Music). A problem of user studies where a searcher is asked to assume someone else's information need is that different searchers might look for different aspects of the information need (Saracevic et al., 1988). This problem should be taken into account in the follow-up study.

Source of variation	Sum of Square	df	Mean of Square	F-ratio
Interface Styles	.37	1	.37	1.572
Query Difficulty	1.425	2	.712	30.267*
Interaction	.066	2	.033	1.394
Within Cell	1.271	54	.024	
Total	2.798	59	.047	

Table 4. Two-way ANOVA analysis of corrected precision

Tukey's HSD follow-up test for query difficulty revealed that participants performed poorly with the difficult information tasks on both interface styles ($F < .05$).

Query Difficulty	Interface Styles			
	BiblioMapper		Kohonen SOM	
	Mean	SD	Mean	SD
Easy	.76	.10	.63	.14
Medium	.52	.15	.56	.18
Difficult	.35	.21	.29	.04

Table 5. Mean and standard deviation for corrected precision.

Navigation

Navigation was observed by tracing the paths users followed through the systems. To gain a general measure of navigation, I calculated the number of documents selected (Table 6). Such data can be interpreted in more than one way. Obviously, users who are in difficulty are likely to view the document surrogates as they seek to locate relevant information, thus high navigation scores would be a sign of the user having a poor sense of the system structure. On the other hand, viewing various surrogates might be seen as a sign of user comfort with exploration, although this interpretation is less common in the hypertext and menu navigation literature (see e.g., Norman, 1991). My observations of the users in this study support the latter interpretation. In the post-interview session, four of the BiblioMapper subjects stated that BM helped them understand the relationships among the sub-disciplines of computer and information science revealed in the test data collections. Analysis indicated that there is not a significant effect for interface style ($p > .05$).

Query Difficulty	Interface Styles			
	BiblioMapper		Kohonen SOM	
	Mean	SD	Mean	SD
Easy	23.50	9.18	25.00	9.67
Medium	27.50	12.72	25.80	6.69
Difficult	30.80	14.84	33.10	8.09

Table 6. Number of documents selected (mean and standard deviation).

Preference ratings

Participants were asked to comment on the interfaces in the search sessions and any difficulty they had in using the interfaces. Six of the participants expressed a preference for the Kohonen SOM interface whereas feelings of control over the system were higher for the BiblioMapper interface.

However, their performance with Kohonen SOM in terms of completion time, accuracy, and the number of document selections was at best equal to BiblioMapper. One interesting comment by subjects who preferred the Kohonen interface was "I don't like the BiblioMapper interface. It took time to search with the BiblioMapper." It was apparent that the subjects who made this comment were not motivated to locate the relevant documents.

One major observation with the Kohonen SOM searching is that participants tended to rely on the alphabetical subject list next to the map rather than the map itself. The map was consulted only when the subjects were willing to spend more time on searching for other subject areas relevant that appeared to be interesting.

Conclusion

It was initially assumed that Kohonen SOM, whose theoretical base follows the associative neural properties of the brain, will correspond to users' mental models more properly than BiblioMapper. This assumption was not substantiated in this experiment.

Obviously the task type is a major source of the variance in user response to interface style, and searches involving keywords which do not appear in the visual space were considerably difficult and took time. This is one major challenge for both interfaces.

This experiment with users' performance on BiblioMapper and Kohonen SOM confirms the findings of the previous study that visualization of document space provided by both BiblioMapper and Kohonen SOM, which does not support querying, was suitable for browsing

searches rather than for analytical searches (Song, 1998). Similar results with WEBSOM, a visualization system of a document space, were reported by Orwig et al. (1997).

According to the results, BiblioMapper and Kohonen SOM may be useful in support of the kind of exploratory activities that occur when users encounter large unknown text collections, but it should be coupled with other kinds of retrieval techniques, such as query-based search, that can be enhanced by the knowledge that users gain through preliminary BiblioMapper browsing.

Since the results reported are based on an exploratory experiment, on-going experiments are needed to verify these results. The follow-up study will focus on investigating whether these information exploration tools can be used effectively and scaleably to satisfy users' information needs which are not well-defined and require "true browsing tasks."

Appendix

The search tasks

1. Easy task:

I want to find out some projects related to multi-cast.

2. Medium task:

As one approach to system design, adaptive system design has been drawn much attention by designers. I need to know what adaptive system design is and in what setting it has been applied.

3. Difficult task:

I am looking for evaluation methods which enable me to identify some design problems and to re-design the system.

References

- Belew, R. W. 1989. Adaptive Information Retrieval. In Proceedings of 12th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval.
- Chen, H., C. Schuffels, and R., Orwig. 1996. Internet Categorization and Search: A Machine Learning Approach. *Journal of Visual Communications and Image Representation*, 7(1): 88-102.
- Chen, H. A. L., Houston, R. R. Sewell, and B. R. Schatz. 1998. Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques. *Journal of the American Society for Information Science*, 49(7): 582-603.
- Golocvchinsky, G. 1997. Queries? Links? Is There a Difference? In Proceedings of CHI '97, 407-414, March.

- Kohonen, T., T. Honkela, S. Kaski and K. Lagus. 1996. Newsgroup Exploration with WEBSOM Method and Browsing Interface. Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science.
- Lin, X., D. Soergel, and G. Marchionini 1991. A Self-Organizing Semantic Map for Information Retrieval. In Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 262-269.
- Meadow, C. 1992. Text Information Retrieval Systems, New York: Academic Press.
- Norman, K. L. 1991. The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface, Norwood, N.J.: Ablex Publishing.
- Orwig, R. E., H. Chen, and J. F. Nunamaker, Jr. 1997. A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output. Journal of the American Society for Information Science, 48(2): 157-170.
- Ritter, H. And T. Kohonen 1989. Self-Organizing Maps. Biological Cybernetics, 61, 241-254.
- Saracevic, T., P. Kantor, A. Y. Chamis, and D. Trivison. 1988. A study of information seeking and retrieving. Journal of the American Society for Information Science, 39 (3), 197-216.
- Song, M. 1998. Can Visualizing Document Space Improve Users' Information Foraging? Accepted for ASIS 1998 Annual Meeting: Information Access in the global Information Economy, October 25-26.
- Song, M. And T. Gillespie 1997. BiblioMapper: Visualizing a Database of Technical Reports. In HCI International '97: 7th International Conference on Human-Computer Interaction Jointly with 13th Symposium on Human Interface (Japan), pp. 87-88, August 24-29.
- Tague-Sutcliffe, J. 1995. Measuring Information: A Information Retrieval System: A Information Services Perspective. San Diego, CA: Academic Press.
- Veerassamy, A. And Heikes R. 1997. Effectiveness of a Graphical Display of Retrieval Results. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 236-245.
- Williams, J. G., K. M. Sochats, and E. Morse 1995. Visualization. Annual Review of Information Science and Technology, vol. 30, pp. 161-207.

