Katy Börner (2000) Visible Threads: A smart VR interface to digital libraries. Proceedings of IST/SPIE's 12th Annual International Symposium: Electronic Imaging 2000, Visual Data Exploration and Analysis (SPIE 2000), San Jose, CA, 23-28 January, pp. 228-237. <http://ella.slis.indiana.edu/~katy/SPIE00/>

Visible Threads: A Smart VR Interface to Digital Libraries

Katy Börner[*]

Indiana University, School of Library and Information Science
10th Street & Jordan Avenue, Bloomington, IN 47405

ABSTRACT

The importance of information as a resource for economic growth and education is steadily increasing. Due to technological advances in computer industry and the explosive growth of the Internet much valuable information will be available in digital libraries. This paper introduces a system that aims to support a user's browsing activities in document sets retrieved from a digital library. Latent Semantic Analysis is applied to extract salient semantic structures and citation patterns of documents stored in a digital library in a computationally expensive batch job. At retrieval time, cluster techniques are used to organize retrieved documents into clusters according to the previously extracted semantic similarities. A modified Boltzman algorithm [1] is employed to spatially organize the resulting clusters and their documents in the form of a three-dimensional information landscape or "i-scape". The i-scape is then displayed for interactive exploration via a multi-modal, virtual reality CAVE interface [8]. Users' browsing activities are recorded and user models are extracted to give newcomers online help based on previous navigation activity as well as to enable experienced users to recognize and exploit past user traces.
In this way, the system provides interactive services to assist users in the spatial navigation, interpretation, and detailed exploration of potentially large document sets matching a query.

Keywords: Digital Libraries, Singular Value Decomposition, Latent Semantic Analysis, Boltzman Algorithm, Conceptual Clustering, Information Visualization

1. INTRODUCTION

The amount of information available in digital libraries is increasing rapidly and at accelerating rate. At the same time the pressures to use information sources efficiently and the diversity of information needs are also increasing. Oftentimes, users can only vaguely specify what they are looking for. Frequently, they start with a general topic of interest, retrieve relevant documents via a search engine, and browse the retrieval results. However, a long list of hits can be overwhelming.

The visualization of retrieval results in the form of *information spaces* or "i-scapes" may help human users to discover semantic relations in data. Given that certain principles of visual presentation are followed [5,18,19,20], i-scapes can provide an immediate overview of an unfamiliar terrain. They are meant to guide one's way and to reduce the time spent wandering in the wilderness.

The paper describes a prototypical system that supports users' retrieval-browsing behavior by visualizing retrieval results for fast navigation and exploration. The system automatically extracts semantic relationships in document sets and groups semantically similar documents into clusters. A modified Boltzman algorithm [1] is used to lay out clusters and documents in a visually acceptable way. The result is displayed for interactive exploration at different levels of detail. The implemented virtual reality interface provides a rich, three-dimensional representation of clusters and documents in the form of an information landscape. It is augmented with navigation capabilities to explore the documents and their relations in detail, thus enabling users to comprehend and interact with the information in an efficient way. Almost no knowledge of the document set prior to browsing is required. The i-scape visualization is designed to help users to understand the retrieved document sets as a whole, and to select and examine relevant documents on the fly. Additionally, user models are extracted from the user's navigation path through documents as well as from modifications of document relevance values. The user models are applied to give newcomers online help based on previous navigation activity and to enable experienced users to recognize and exploit past user traces. Figure 1 (which we discuss in a later section) shows a snapshot of the interface.

In the following section we provide an overview of related work. We then introduce our approach to data analysis and information visualization, describe its implementation, and present a sample session with the system. The paper concludes with a discussion of the results and an outlook.

[*] Correspondence: Email: katy@indiana.edu; WWW: http://ella.slis.indiana.edu/~katy; Telephone: 812 855 3256;
Fax: 812 855 6166

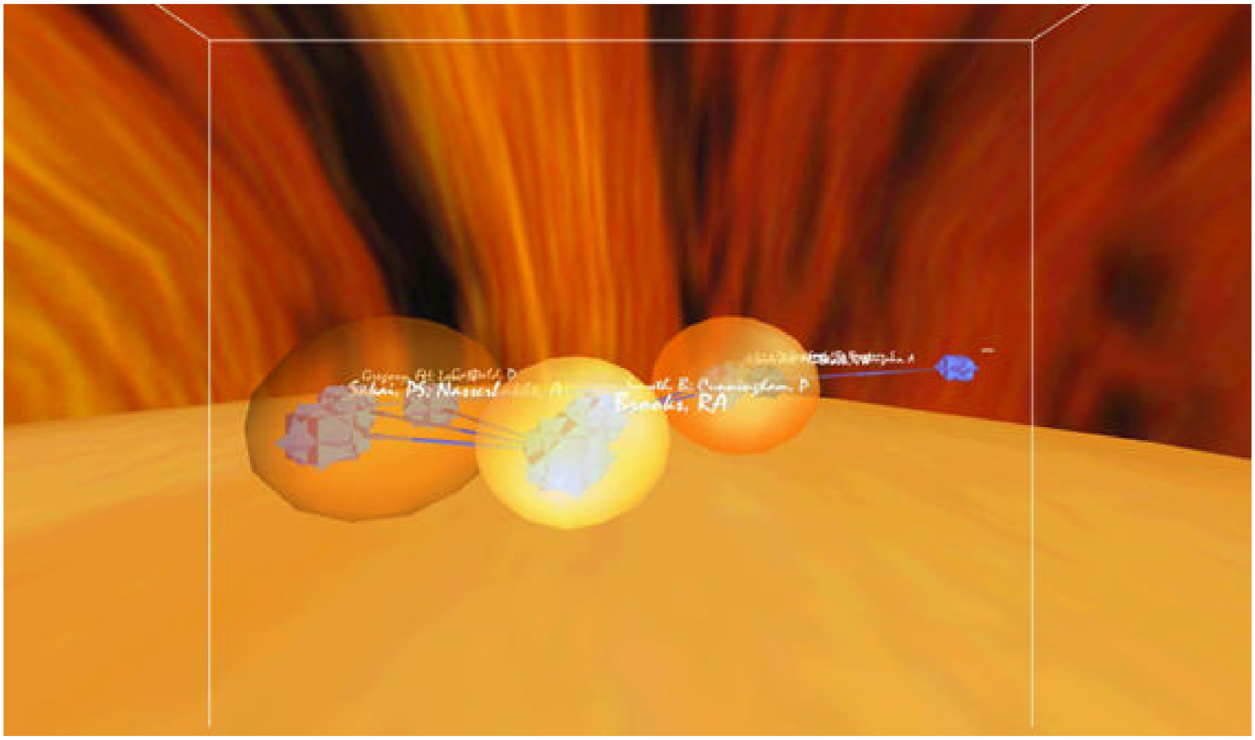

Figure 1: CAVE Interface showing Document Clusters

2. RELATED WORK TO DATA ANALYSIS AND VISUALIZATION

Most digital libraries provide search engines to access documents. Retrieval results are usually displayed as a linear, sorted list of matching documents. Users often browse through the first (Web) page of documents and skip the remaining hits, overwhelmed from the high number of matches.

Subsequently, we discuss approaches and systems that use statistical techniques to automatically extract semantic patterns from text document sets. The resulting semantic relationships are then used to offer a visual guide to organize huge amounts of data, to make it easy to find relevant documents, and to provide a quick overview of surrounding documents. Some systems display thousands of documents - often grouped into topics - simultaneously. Given the lack of a natural spatial organization of most document sets, a variety of metaphors can be applied to generate meaningful visualizations and interactions that match the spatial abilities of their users [2]. "Star field" and "landscape" metaphors are used frequently to organize documents in a way that is natural for the human mind and requires little training to use. Both metaphors allow spatial models of interaction to be used for the navigation and the manipulation of documents.

Research by Chen aims to support people's exploration and navigation activities in a semantically organized information space [6]. As a sample document set, 169 long papers from three years ACM SIGCHI Conference Proceedings (1995-1997) and all papers from the ACM Hypertext Conference Proceedings (1987-1998) were selected. A document-document similarity matrix was generated via Latent Semantic Analysis (explained in the next section). Each document was indexed by its title, authors' affiliations, abstract, and list of keywords. *Pathfinder Network Scaling* was used to extract underlying patterns in the similarity matrix and to present them spatially in a class of 'pathfinder networks' [15]. The approach was implemented in StarWalker (see http://www.brunel.ac.uk/~cssrccc2/vrml2/starwalker/ccpro.htm) a system that displays several precompiled citation networks. Visually, important papers (according to a citation threshold) are represented by spheres. Links between two spheres indicate that those two authors have been cited together to an extent determined by the author citation analysis. Clicking on a sphere results in the download and display of the original full text of its article at ACM's electronic proceedings website for detailed study [7].

Researchers at the Pacific Northwest National Laboratory have developed a suite of new technologies called SPIRE™, or Spatial Paradigm for Information Retrieval and Exploration [22]. SPIRE accepts large volumes of unformatted text, determines the dominant topics and relationships within the text, and graphically displays them based on word similarities and themes in text such that similar texts are constrained to lie close to one another. The visualizations allow the user to explore and rapidly discover known and hidden relationships among text documents and to quickly extract and use the desired information. The research aims at a multi-faceted insight and understanding of data through interoperable visual paradigms [12]. Galaxies, Themeview, and Tumbleweed are technologies within SPIRE.

Galaxies computes word similarities and patterns in documents. It then displays the documents on a computer screen to look like a universe of "docustars." Closely related documents will cluster together in a group and unrelated documents will be separated by large spaces [21].

Themeview uses *Multidimensional Scaling* to reduce the document vector from an n-dimensional space into two-dimensional space that is used for visualization in the form of a relief map of natural terrain. Mountains indicate dominant themes; valleys indicate weak themes. Their shapes -- a broad butte or high pinnacle -- reflect how the thematic information is distributed and related across documents. Themes close in content are close visually, where content is based on the many relationships within the text spaces [21].

Tumbleweed uses *Latent Semantic Analysis* to determine context vectors for documents. Context vectors for topics are obtained as a natural adjunct of the document context vectors. *Principal component analysis* is used to transform topic vectors from n-dimensional coordinates to normalized 3-dimensional coordinates. The three most salient topics are used as axes and documents are positioned in space according to the two or three most closely related topics (determined via Euclidean distance). This way particular locations in the created topical layout have an inherent meaning.

The SPIRE technology is used by NewsMaps, from Cartia, Inc. to index the contents of documents such as Web sites, newsgroups, or other database entries by word co-occurrences. Its user interface at http://www.newsmaps.com looks like a topographical map, with 'hills' and 'valleys' representing clusters of related documents. Clicking on a mountain pops up a menu with a list of topics. By scrolling through topics inside the menu box, users can find the document they have been looking for or they can click on the topic and all related documents are highlighted in the map. If users click on a document, NewsMap launches a separate window with the requested page. To keep track of their history of visited sites, users can plant flags. Because of the expense of data analysis and map generation done behind the scene, NewsMaps is only updated two or three times a day.

Work by Small aims to generate a map of the "World of Science" [17] based on highly cited papers and their frequency of co-citation. He examined 36,720 multidisciplinary documents for a 15-year sampling window from 1981 through 1995. A combination of fractional citation counting[1] and co-citation clustering via *Multidimensional Scaling* was used to extract four nested levels of clusters via single and complete linkage. The documents are arranged in 2D using an order-dependent, geometric triangulation process that produces a unified ordination of a hierarchical arrangement of documents. The resulting "Map of Science" [17,10] is a series of nested maps showing the multi-dimensional landscape of science at five levels of aggregation. In these maps, each large area of research is represented by a circle. The size of the circle is proportional to the number of papers published. The distance between the centers of these circles is related to the number of co-citations between the fields. The structure of the maps shows how disciplines, fields, specialties, and individual papers or authors relate to one another.

All systems mentioned so far determine the semantic structure or co-citation networks as well as the final layout of predefined document sets in a computationally expensive process. The resulting visualizations are mostly static. Narcissus is the name of a sophisticated visualization tool for web pages that uses self-organizing systems and a modified version of XMosaic to generate three-dimensional information spaces of 'active objects' that a user can navigate and manipulate [14]. All active objects follow the same set of task-dependent rules: they exert a repulsive force on all of the other objects and they attract objects that are related. Users determine by hand which relationships are active within the model, thus influencing the final layout. Running the model, the objects successively arrange themselves in space close to objects to which they are related. A stable state can still include some pulsating motion, adding to the 3-D effect and the recognition of similar patterns. However, for larger data sets the computational cost of determining object positions and the time required to reach a stable state may be too high. The specification of active relationships between objects is task- and data set-dependent and requires specific knowledge.

[1] Each citation is inversely weighted by the length of the reference list of the citing article.

The research described in this paper applies Latent Semantic Analysis to automatically extract the semantic structure of a particular digital library in a computationally expensive batch job. At retrieval time, the result of a query is organized using hierarchical clustering techniques and visualized for browsing according to the extracted underlying structure. User models are extracted to provide additional guidance.
While all of the reviewed systems use a desktop interface and 2-D devices such as a keyboard and a mouse for navigation and manipulation of documents, our system provides a virtual reality CAVE interface that merges visual and manipulative spaces, immersing a user in the data.

3. DATA ANALYSIS

In order to identify semantic relationships of a set of documents for display, we use Latent Semantic Analysis (LSA) to extract document-document similarity matrices in a computationally expensive batch job. At retrieval time, the document-document similarity matrices for different terms such as author name(s), document title, abstract, keywords, or cited references allow documents to be organized according to these terms or combinations thereof. Clustering techniques are applied to determine the best partitioning of retrieved documents into clusters.

3.1. Latent Semantic Analysis

Latent Semantic Analysis is an extension of the vector space model for information retrieval [9,4,13]. In the vector space model, a representative sample of documents is represented by a term-document matrix X in which each cell indicates the frequency with which each term (row) occurs in each document (column). Consequently, a document becomes a column vector and can be compared with a user's query represented as a vector of the same dimension. The similarity between a query vector and a document vector is usually measured by the cosine of the angle between them. In such a way, a list of documents can be determined, ranked according to their similarity to the query, and returned to the user. LSA extends this model by modeling the term-document relationships by a reduced rank-k approximation of the original term-document matrix computed via Singular Value Decomposition (SVD).

SVD decomposes a rectangular term-document matrix X into the product of three other matrices W, S, and P' with {X} = {W} {S} {P}'. W is an orthonormal matrix and its rows correspond to the rows of X, but it has m columns corresponding to new, specially derived variables such that there is no correlation between any two columns; i.e., each is linearly independent of the others. P is an orthonormal matrix and has columns corresponding to the original columns but m rows composed of derived singular vectors. The third matrix S is an m by m diagonal matrix with non-zero entries (called singular values) only along one central diagonal. A large singular value indicates a large effect of this dimension on the summed-squared error of the approximation. The role of these singular values is to relate the scale of the factors in the other two matrices to each other such that when the three components are matrix multiplied, the original matrix is reconstructed.
The implemented system uses a sparse SVD via single-vector Lanczos algorithm from M. Berry's SVDPACK [3] for singular value decomposition, because of its comparatively low memory and processing requirements while maintaining acceptable accuracy.

Following the decomposition by SVD, the k most important dimensions (those with the highest singular values in S) are selected. All other factors are omitted, i.e., the other singular values in the diagonal matrix along with the corresponding singular vectors of the other two matrices are deleted. The reduced dimensionality solution then generates a vector of k real values to represent each document. Each of the k dimensions is associated with a pseudo-concept which may not have any explicit semantic label, yet discriminates documents. The reduced matrix ideally represents the important and reliable dimensions representing the data in X.
The amount of dimensionality reduction, i.e. the choice of k, is critical and an open issue in the factor analytic literature. Ideally, k should be large enough to fit the real structure in the data, but small enough such that noise, sampling errors or unimportant details are not modeled [9]. We use good retrieval and browsing performance as an operational criterion.

3.2. Mathematical Clustering Techniques

At retrieval time, mathematical clustering techniques are used to group (and later on display) semantically closely related documents into a set of clusters. Specifically, nearest-neighbor-based, agglomerative, hierarchical, unsupervised conceptual clustering is applied to create a hierarchy of clusters grouping documents of similar semantic structure. Clustering starts with a set of singleton clusters, each containing a single document $d_i \in D$, i=1, ..., N, where D equals the entire set of documents and N equals the number of all documents. The two most similar clusters over the entire set D are merged to form a new cluster that covers both. This process is repeated for each of the remaining N-1 documents. An average linkage algorithm is applied to determine the overall similarity of document clusters based on the document-document similarity matrix as determined via LSA. Merging of document clusters continues until a single, all-inclusive cluster remains. At termination, a uniform, binary hierarchy of document clusters is produced.

The partition showing the highest within-cluster similarity and the lowest between-cluster similarity is used for data visualization. It is determined by means of a *utility* measure that contrasts the sum of within-cluster similarities $wSim$ by the sum of between-cluster similarities $bSim$:

$$utility = \frac{wSim}{wSim+bSim}.$$

Each displayed cluster is labeled by the keyword that occurs most often in its documents to quickly convey information about the topics by which documents are grouped. Together with their documents, clusters compose a dynamic visual i-scape with which the user can interact to explore the content of a retrieval result.

4. INFORMATION VISUALIZATION

An adequate visualization of retrieval results should make it easy to efficiently and effectively identify relevant documents as well as to gain an overview about the visible threads that exist among documents according to their intrinsic semantic relations. Given the potentially large number of documents matching a query, overview, zoom, filter, and "details on demand" facilities [16] are important. These techniques also help one to handle the universal challenges of time and computation tradeoffs connected with the visual display of large document sets.

The implemented interface gives users access to three levels of detail:

- It provides an overview about document clusters and their relations.
- It shows how documents belonging in the same cluster relate to one another.
- It gives detailed information about a document such as author(s) name, document title and source, cited references, abstract, language, document type, keywords, and times cited.

Rather than being a static visualization of data, a modified Boltzman algorithm [1] is used to provide a visualization that is self-organizing and highly interactive. The algorithm works by computing attraction and repulsion forces among nodes. In our application, the nodes represent documents retrieved by a search engine. They are attracted to other nodes to which they have a (reference or similarity) link and repelled by nodes to which there is no link. Attraction and repulsion forces are computed based on the underlying document-document similarity matrix. The algorithm starts out with a random arrangement of nodes. Over several iterations, nodes that are similar according to the document-document matrix are pulled together and nodes that are dissimilar are pushed apart. The system usually stabilizes after a few seconds resulting in a more-or-less acceptable layout of data that reflects the underlying semantic relationships. The algorithm is applied at two levels: the level of clusters and the level of documents in each cluster. In order to guarantee a deterministic layout, clusters (as well as documents in clusters) are placed in space alphabetically according to their cluster label (documents are placed according to the name of their first author). This way, layouts of the same query result resemble each other, making it easy for users to orient themselves in the i-scape. Different query result sets, however, can produce entirely different layouts. Resulting positions of clusters and documents can be modified by a user as desired.
Different document-document matrices (e.g. for keywords, abstracts etc) result in different layouts enabling the user to see the underlying i-scape from different perspectives and thus helping to better understand the information available.

5. CAVE INTERFACE

A good visualization supports three user tasks: It communicates what is already known; it confirms or refutes hypotheses; and it supports exploratory browsing [11]. This section introduces the system's interface and a walk through of a simple sample information-retrieval-browsing scenario.

In order to design an interface that conforms to the expectations of everyday reality, the CAVE system [8] at Indiana University's Advanced Visualization Laboratory was used. It enables one to create a multi-modal, virtual environment in which sets of document retrieved via the Web of Science are mapped onto an i-scape. The landscape is intended to give the user an immediate overview about the relationships between the documents, the times a particular document is cited, and the importance other users gave to a document. This is accomplished by linking the objects of interaction tightly to the usage information.

5.1. Data Retrieval and Analysis

The current prototype of the system displays retrieval results from the ISI® Web of Science® Interface at http://webofscience.com/ . Querying a citation database such as the Science Citation Index Expanded via the Web of Science® Interface at http://webofscience.com/ results in a maximum number of 500 document references organized in lists of ten. Retrieved documents can be marked, saved, and downloaded for detailed study. The documents are represented in the usual Web of Science data output format (including author(s), document title and source, cited references, addresses, abstract, language, publisher information, ISSN, document type, keywords, times cited, etc.) as shown in Figure 2.

Latent Semantic Analysis was applied over keywords, abstracts and co-citation links of a large document set from the Science Citation Index Expanded. The resulting document-document similarity matrices act as input to the visualization routines.

```
FN ISI Export Format
AU Brooks, RA
TI From earwigs to humans
SO ROBOTICS AND AUTONOMOUS SYSTEMS
DT Article
PU ELSEVIER SCIENCE BV
DE humanoid robotics
ID NEURAL NETWORKS; BEHAVIOR; SYSTEMS; ROBOT; MODEL
AB Both direct, and evolved, behavior-based approaches to mobile
   robots have yielded a number of interesting demonstrations robots
   that navigate, map, plan and operate in the real world.  ...
CR AGRE PE, 1987, P AAAI 87, P268
   ARKIN RC, 1992, ADAPT BEHAV, V1, P201
   ASHBY WR, 1952, DESIGN BRAIN
   ...
TC 3
JI Robot.  Auton.  Syst.
PY 1997
```

Figure 2: Partial ISI Document Representation

5.2. Data Visualization

Users can start the CAVE application with three parameters that specify the

- The retrieval result to be displayed.
- The viewpoint under which retrieved documents should be organized (similar keywords, abstracts, co-citation links or combinations thereof).
- A similarity threshold determining how closely related documents have to be to a query in order to be displayed.

Figure 1 shows the CAVE interface to a retrieved document set. Each document corresponds to a node in a dependency graph. The shape of a node represents the type of the document (journal article, book article etc.). Node color corresponds to the year of publication of a document. Its label identifies the first author. Incoming links represent a document's references. Outgoing links denote the document's citers.

The x and y dimension are used to place documents in space such that semantically similar documents are close to each other. The z dimension is used to reflect different views of documents such as the 'relevance of documents' according to citation counts - frequently cited documents rise to the top, like cream. The importance of a document as assigned by other users is reflected by the size of the node. Semantically similar documents are grouped into clusters. Documents belonging in the same cluster are enclosed by translucent spheres that are labeled by the most frequent keyword occurring within the sphere's documents. The higher the similarity of the documents within a cluster, the lighter is the color of the corresponding sphere. Table 1 provides an overview of the used visualization model.

A strength of the interface comes from its ability to flexibly interact with a set of documents. The 'cluster view' provides an immediate overview of the different topics into which documents can be grouped and the number of co-citation links among documents belonging to different clusters. The 'document view' reveals details about its documents, their citation links, and the importance other users gave to documents. Citation data of each displayed document can be displayed in a Web browser interface.

digital object	geometric model	attribute	semantics
document	crystal	color	year of publication
		shape	document type
		size	relevance
		x-y-position	depends on similarity relations to other documents
		z-position	times cited
	label		first author name
document link	arrow		citation link or similarity link
cluster	sphere	size	number of documents
		color	within-cluster similarity
		x-y-position	depends on similarity between clusters
	label		most frequent keyword of its documents
cluster link	arrow		citation link
		width	number of citation links

Table1: Visualization Model

Starting with a particular document set and viewpoint, the user can interact with the i-scape in various ways using a 3D input device called a wand. The wand has three buttons and a pressure-sensitive joystick. It is connected to an Ascension Flock of Birds tracking system. The joystick allows the user to spatially navigate in the world of displayed clusters, documents, and co-citation links. If the wand is close enough to a document, it is highlighted indicating a successful selection. Users can pick and move a selected document to change the layout of documents. A subsequent push of the same button releases the document at the current wand position. Selection of a document and pushing the middle button results in the display of a web browser window (see Figure 3, left) that provides details about the document. Simple audio cues are used to provide feedback about object manipulations.

Figure 3: CAVE - Display of Citation Data

By moving documents into a "personal basket", users can retrieve detailed information about documents via mail at the end of a session. The CAVE interface is mainly used to gain an overview about key topics; to see documents related to those known to the user, and to select documents. E-mail containing detailed citation data is provided for extended, focused reading.

In addition to this, the user can change the size of a selected node corresponding to the importance s/he assigns to the corresponding document by pushing the left wand button. Increasing the size of a node also facilitates finding documents related to it while navigating through the space of documents. At the end of a session the changed sizes assigned to particular nodes are written out as guidance for subsequent users.

As information spaces become more complex, the derivation and exploitation of user models becomes increasingly valuable for solving information navigation tasks "collaboratively". They can provide newcomers with online help based on the previous activities of other users as well as enable experienced users to recognize and exploit past user traces. Virtual reality environments are shared by humans and computers and permit tracking of human behavior in an extensive way. In the present implementation, two types of user interaction (which we term *passive* and *active*) are recorded as a basis for subsequent support:

- The user's navigation path through documents (*passive* - like the "history" or "go-to" list in a Web browser).
- The modification of node sizes (*active* - similar to a list of user-created "bookmarks").

Although the relevance of documents may decrease (or increase) over time, we do not deal with these fluctuations currently. The issue of privacy, though important, is not relevant to our research. All participant results are anonymous and are compiled into a common database.

To demonstrate a visual browser to the Web of Science® interface we use CogSci200, a query result data set from the Science Citation Index Expanded that contains the 200 most cited documents matching the key 'Cognitive Science'. Launching the CAVE application with this document set, setting abstracts as the viewpoint[2], and setting a similarity threshold of 0.5, an initial I-scape as shown partially in Figure 1 is displayed in the CAVE.
An imaginative user interested in an overview about influential articles in cognitive science may examine the clusters to get an overview about current topics, s/he may navigate to relevant clusters to examine the similarity relations among their documents based on dominant word co-occurrences in their abstracts, or have a look at the existing co-citation patterns. In Figure 3 an article by *Rodney A. Brooks* was selected and its citation data was retrieved and displayed in a Web browser on the right hand side. By moving the document's node into the personal basket, the user will receive an email containing the citation data at the end of the session.

Colored, full-size versions of Figures 1 and 3 are accessible at http://ella.slis.indiana.edu/~katy/SPIE00.

6. CONCLUSIONS AND FUTURE WORK

This paper introduced an approach to the automatic extraction and 3-D visualization of document retrieval results for browsing. A sample implementation of the approach was presented together with a sample usage scenario. The main novelty of the presented research is the application of LSA over a database in a preprocessing step. The derived document-document similarity matrices are then used at retrieval time to display matching documents -- a subset of the database -- in the CAVE for visual browsing.
The approach and its implementations are in early development. An obvious next step is to evaluate the merit of the interface and the interaction with the data it provides, and to test the scalability of the approach on larger datasets. Larger data sets will increase the need for hierarchical organizations of documents and a facility to zoom into clusters. A picked cluster could expand unraveling its documents and their relations, accompanied by a shrinking of all other clusters. The interactive change between different viewpoints (instead of launching the application with one particular viewpoint) is under development as well. The usefulness of the clusters as extracted by LSA is very user- and context-dependent (i.e., it relates to the original query). Therefore, we plan to use filtering techniques to extract user (group) profiles. These profiles will be used to personalize the information visualization such that users are guided in their exploration of various data relationships according to their individual interests and needs.

In general, powerful mathematics applied to identify semantic relations in document sets combined with information visualization techniques may enable systems to be built that can help users explore, understand, and remember complex

[2] Even though the system offers the facility to run LSA over citation links, this parameter is rarely used because of the very few citation links (often referring to works by the same author) in the document sets retrieved.

relationships in document sets, and to select the set of documents relevant for them. The development of innovative interfaces to digital libraries has only begun. Their integration with existing technology will improve the way we understand and access the enormous amounts of data available today.

ACKNOWLEDGEMENTS

The author wishes to thank Indiana University's Advanced Visualization Laboratory for providing resources and assistance. The SVDPACK [3] was used for computing the singular value decomposition of large sparse matrices.
Special thanks go to Robert Goldstone for insightful comments and suggestions. He and Mark Steyvers have been valuable discussion partners and influenced the selection and optimization of the LSA algorithm applied.
The research is supported by a *High Performance Network Applications* grant of Indiana University. Collaborators in this project are Andrew Dillon, Margaret Dolinsky, and Tyler Waite.

REFERENCES

1. Alexander, Garcia, & Alder. Simulation of the Consistent Boltzman Equation for Hard Spheres and Its Extension to Dense Gases, *Lecture Notes in Physics*, Springer Verlag, 1995.

2. Allen, B., "Information Space Representation in Interactive Systems: Relationship to Spatial Abilities", Proceedings of the third ACM Conference on Digital Libraries, Pittsburgh, pages 1 – 10, 1998.

3. Berry, M., Do, Th., O'Brian, G. W., Krishna, V., & Varadhan, S., "SVDPACKC (Version 1.0) User's Guide," *University of Tennessee Tech. Report CS-93-194*, 1993 (Revised October 1996). See also http://www.netlib.org/svdpack/index.html.

4. Berry, M, Dumais, S. T. & O'Brien, G. W., "Using linear algebra for intelligent information retrieval," *SIAM: Review*, 37(4), 573-595, 1995.

5. Brath, R., "Concept demonstration: Metrics for effective information visualization," In *Proceedings for IEEE Symposium on Information Visualization*, pages 108-111. IEEE Service Center, Phoenix, AZ, Oct 20-21, 1997.

6. Chen C., "Visualizing semantic spaces and author co-citation networks in digital libraries," *Information Processing and Management* **35**, 401-420, 1999.

7. Chen, C., & Carr, L., "Trailblazing the literature of hypertext: Author co-citation analysis (1989-1998)," *Proceedings of the 10th ACM Conference on Hypertext*, 1999.

8. Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. "Surround-screen projection-based virtual reality: The design and implementation of the CAVE", in J. T. Kajiya (ed.), *Computer Graphics* (Proceedings of SIGGRAPH 93), Vol. **27**, Springer Verlag, pp. 135-142, 1993.

9. Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A., "Indexing By Latent Semantic Analysis," *Journal of the American Society For Information Science* **41**, 391-407, 1990.

10. Garfield, E., "Mapping the World of Science," Presentation - Topical paper presented the *150th Anniversary Meeting of the AAAS*, Philadelphia, PA. February 14, pp 1-19, 1998, http://www.garfield.library.upenn.edu/papers/mapsciworld.html.

11. Grinstein, G. & Ward, M., "Introduction to Data Visualization," *IEEE Visualization'97*. Phoenix AZ, October 1997.

12. Hetzler, B., Whitney, P., Martucci, L., & Thomas, J., "Multi-faceted Insight Through Interoperable Visual Information Analysis Paradigms," In *Proceedings of IEEE Symposium on Information Visualization*, October 19-20, Research Triangle Park, North Carolina. pp. 137-144, 1998.

13. Landauer, T. K., Foltz, P. W., & Laham, D., "Introduction to Latent Semantic Analysis," *Discourse Processes* **25**, 259-284, 1998.

14. Hendley, R. J., Drew, N. S., Wood, A. M., & Beale, R., "Narcissus: Visualising Information," *Proceedings of IEEE Symposium on Information Visualisation*, Atlanta, USA. 1995.

15. Schvanenveldt, R. W., Durso, F.T, and Dearholt, D. W., "Network Structures in Proximity Data," *The Psychology of Learning and Motivation*, G. Brower, ed, Vol. **24**, Academic Press, San Diego, California, pp. 249-284, 1998.

16. Shneiderman, B., "The Eyes have it: A task by data type taxonomy for information visualization". In *Proceedings for IEEE Symposium on Visual Languages*, pp. 336-343, Sept. 3-6, IEEE, 1996.

17. Small, H., "Visualizing science by citation mapping," *Journal of the American Society for Information Science*. **50**(9):799-813, 1999.

18. Tufte, E. *Envisioning Information*. Graphic Press, 1990.

19. Tufte, E. *The Visual Display of Quantitative Information*. Graphic Press, 1993.

20. Tufte, E. *Visual Explanations*. Graphic Press, 1997.

21. "Visual Text Analysis: SPIRE. Visual Text Analysis – SPIRE," A technical flier from the Pacific Northwest National Laboratory," http://multimedia.pnl.gov:2080/showcase/pachelbel.cgi?it_content/spire.

22. Wise, J. & Thomas, J., "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents," In *Proceedings of IEEE Information Visualization*, pages 51-58. IEEE Service Center, Atlanta, GA, October 1995.