Börner, Katy. (2000). Extracting and visualizing semantic structures in retrieval results for browsing. ACM Digital Library Conference, San Antonio, Texas, June 2-7, pp. 234-235.

Extracting and Visualizing Semantic Structures in Retrieval Results for Browsing

Katy Börner
Indiana University, School of Library and Information Science
10th Street & Jordan Avenue, Main Library 019, Bloomington, IN. 47405 USA
E-mail: katy@indiana.edu

ABSTRACT
The paper introduces an approach that organizes retrieval results semantically and displays them spatially for browsing. Latent Semantic Analysis as well as cluster techniques are applied for semantic data analysis. A modified Boltzman algorithm is used to layout documents in a two-dimensional space for interactive exploration. The approach was implemented to visualize retrieval results from two different databases: the *Science Citation Index Expanded* and the *Dido Image Bank*.

KEYWORDS: Digital Libraries, Browsing, LSA, Conceptual Clustering, Boltzman Algorithm, Information Visualization

INTRODUCTION
The wealth of digitally stored data available today increases the demand to provide effective tools to retrieve and manage relevant data. Keyword searches over digital libraries, repositories, or the Web easily retrieve lists of several hundreds of documents.
Information visualization - the process of analyzing and transforming data into an effective visual form - is believed to improve our interaction with large volumes of data.
First visual interfaces to digital libraries provided full-text searching and full-content retrieval capabilities and visualized documents according to authors, time, place, or citation relationships.
A considerable body of recent research applies powerful mathematical techniques such as *Factor Analysis*, *Multidimensional Scaling,* or *Latent Semantic Analysis* to extract for example the underlying semantic structure of documents, the (evolving) specialty structure of a discipline, author co-citation patterns, changes in authors' influences in a particular field. In order to display the results of the data analysis spatially, computationally expensive techniques have to be applied to transform data analysis results to 2 or 3-dimensional coordinates. The computational expense of data analysis and visualization generation is very high. Therefore, precompiled, mostly static visualizations of fixed data sets are only displayed.

To our knowledge there exists no system that interactively visualizes retrieval results for browsing based on their underlying semantic structure.

DATA ANALYSIS
Latent Semantic Analysis (LSA) [4] has demonstrated improved performance over the traditional vector space techniques. It overcomes the problems of *synonymy* (variability in human word choice) and *polysemy* (same word has often different meanings) by automatically organizing documents into a semantic structure more appropriate for information retrieval. We apply LSA to extract the semantic structure of a particular database in a computationally expensive batch job.
At retrieval time, the result of a database query is hierarchically organized, based on the LSA output. 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. The two clusters most similar are merged to form a new cluster that covers both. This process is repeated for each of the remaining clusters. At termination, a uniform, binary hierarchy of document clusters is produced. The partition showing the highest within-cluster similarity and lowest between-cluster similarity is selected for data visualization.

DATA VISUALIZATION
Rather than being a static visualization of data, the interface is self-organizing and highly interactive. Data is displayed in an initially random configuration, which sorts itself out into a more-or-less acceptable display via a modified Boltzman algorithm [1]. The algorithm works by computing attraction and repulsion forces among nodes based on the result of the data analysis. Nodes may represent articles or images, which are attracted to other nodes to which they have a (reference or similarity) link and repelled by nodes to which there is no link. If the algorithm does not produce a visually acceptable layout, or if the user wishes to view the results differently, nodes can be grabbed and moved.

PROTOTYPE SYSTEMS
Two systems have been implemented in Java using the data organization and visualization methods described above.

SCI-E: The first system visualizes query results from the Science Citation Index Expanded (TM) as published by the Institute for Scientific Information®. The Citation Index provides access to current bibliographic information and cited references in more than 5,600 journals. Querying it via the Web of Science® Interface at http://webofscience.com/ results in an often huge number of

matching documents organized in lists of ten that can be marked, saved, and downloaded for detailed study.

To demonstrate a visual browser to this kind of data base we will use DAIV188, a query result data set from SCI-EXPANDED that contains 188 articles matching the topic 'data AND analysis AND information AND visualization'.

The articles are represented in the usual Web of Science data output format (including author(s), article title and source, cited references, addresses, abstract, language, publisher information, ISSN, document type, keywords, times cited, etc.).

LSA was applied over keywords and abstracts of articles. As a result of conceptual clustering, the 167th partition was selected for visualization. It contains 20 clusters grouping 1 – 53 articles. Figure 1 shows the Java interface. Each book article is represented by a rectangle and each journal article by an oval. Articles are labeled by their first author. Lines between nodes visually represent co-citation links.

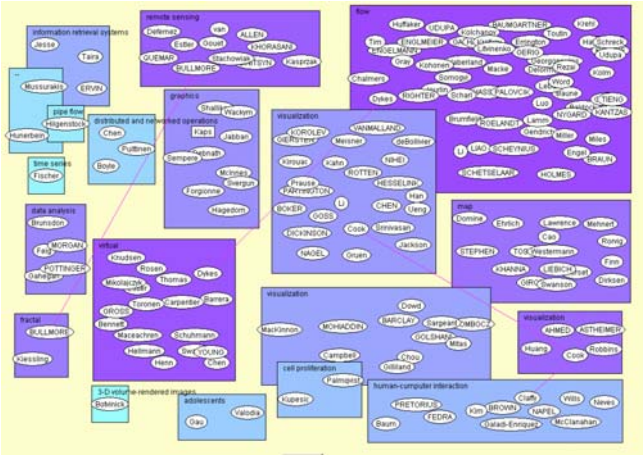

Figure 1: Java Interface to DAIV188

The 2-dimensional layout of articles corresponds to the data mining result as well as to the forces applied by the Boltzman algorithm to generate an acceptable layout. The higher the similarity of articles within a cluster the lighter its color. Each cluster is labeled by the keyword used most often.

DIDO: Another instantiation of the system enables users to browse search results from the Dido Image Bank, http://www.dlib.indiana.edu/collections/dido/ provided by the Department of the History of Art, Indiana University. Dido stores about 9,500 digitized images from the Fine Arts Slide Library collection of over 320,000 images. Each image in Dido is stored together with its thumbnail representation as well as a textual description. LSA was applied over the textual descriptions exclusively. For demonstration purposes the set of images matching the keyword descriptor 'MONET' were retrieved and displayed for browsing. It contains 21 documents inclusive two portraits of Claude Monet drawn by Edouard Manet (see Figure 2).

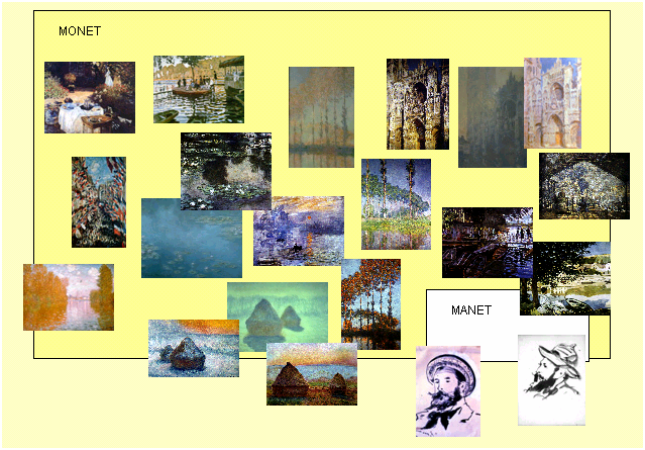

Figure 2: The MONET Cluster

Thumbnail representations of images have been fetched from the Dido Database showing some of Monet's favorite themes such as haystacks, cathedrals, and water lilies.

CONCLUSIONS

Initial tests show that the presented approach provides easy access to textual materials, such as articles, as well as to documents for which textual descriptions are available, such as images. Detailed user studies are in preparation.

First results on using an immersive 3-dimensional CAVE environment for the interactive exploration of search results are presented in [3].

An extended version of this paper as well as colored, full-size versions of Figures 1 and 2 are accessible at http://ella.slis.indiana.edu/~katy/DL00.

ACKNOWLEDGMENTS

Robert Goldstone, Mark Steyvers, Helen Atkins, and Eileen Fry have been valuable discussion partners. The SVDPACK [2] by M. Berry was used for computing the singular value decomposition. The research is supported by an High Performance Network Applications grant of IU. Collaborators are Andrew Dillon and Margaret Dolinsky.

REFERENCES

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

2. Berry, M. et al. SVDPACKC (Version 1.0) User's Guide, University of Tennessee Tech. Report CS-93-194, 1993 (Revised October 1996).

3. Börner, K. Visible Threads: A smart VR interface to digital libraries. *Electronic Imaging 2000, Visual Data Exploration and Analysis*.

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