Towards an Infrastructure for Large-Scale Information Analysis, Visualization, Information Retrieval Research & Education

Katy Börner & Javed Mostafa
School of Library and Information Science
School of Informatics
Indiana University
10th Street & Jordan Avenue, Bloomington, IN 47405
katy@indiana.edu & jm@indiana.edu

Abstract submission for I-Light Conference, University Place Conference Center, Indianapolis, IN, December 4th, 2002.

This research in progress aims to develop an advanced infrastructure for teaching and research in digital libraries, information retrieval, data mining/analysis, and information visualization. The infrastructure will comprise an Oracle database of about 15 million records. We define the term *records* broadly -- covering books, journals, proceedings, doctoral and master's theses, technical reports, patents, grants, and protein and gene data from cross-disciplinary research[1] and specific knowledge domains[2]. Bearing in mind that the ACM portal, CiteSeer, and PubMED currently provide access to about 361,400, 507,800, and 11 million records respectively; this database is unique in its size and coverage. Most of the documents will be available in full text. Software that facilitates a continuous, automatic update of the database will be in place.

An open source software repository[3] [4] will provide access to services such as utility programs (filtering, stemming, stop-word elimination, unique term extraction, automatic citation indexing, etc.), data analysis and dimensionality reduction (e.g., Vector Space Model [11], Multidimensional Scaling [8, 14], Pathfinder Networks [13], Latent Semantic Analysis (LSA) [5, 10], Clustering algorithms, etc.), and visualization/interaction algorithms (GRIDL GRaphical Interface for Digital Libraries [16], Treemap [15], Force Directed Placement [1], and Hyperbolic Tree [9], Self-Organizing Map [7], Fisheye Views [6, 12], the Jazz Zooming Graphics Toolkit [3], etc.). An interchange format based on metadata standards will be developed to ensure that algorithms can be combined in multiple ways (e.g., using different data mining algorithms with diverse information visualization algorithms) [2]. All Java-based algorithms can be run in stand-alone mode as an applet or application. A standardized software framework will interlink the network of different databases and services by a common communication protocol.

All services will run on IU's Sun E10000 Research System (Solar), a shared memory, multiprocessor system with 64 400MHz CPUs and 64GB memory. They can be contributed or requested via the remote graphical user interface (GUI). The infrastructure will directly support the research of the Information Processing Laboratory at IUB. Collaborations with researchers at

[1] Cross-disciplinary DL's comprise ISI's Journal Citation Reports (http://jcrweb.com/), the Proceedings of the National Academy of Sciences (http://www.pnas.org/), Science Magazine (http://www.sciencemag.org), plus awarded NSF (http://www.nsf.gov/) and NIH (http://crisp.cit.nih.gov) grants together with funding opportunities published by the Community of Science (http://www.cos.com/).

[2] Domain specific DLs cover Computer Science (ACM Portal (http://portal.acm.org/), IEEE Xplore (http://ieeexplore.ieee.org/), Networked Computer Science Technical Reference Library (http://www.ncstrl.org/), NEC's CiteSeer (http://citeseer.nj.nec.com/cs), E-Print archive (http://arxiv.org/)), Physics, Mathematics, Nonlinear Sciences (E-Print archive), and Medicine (Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=OMIM)).

[3] http://ella.slis.indiana.edu/~katy/L697/code/

the School of Informatics at IU Bloomington are currently underway to share the resources of the infrastructure across campus. In addition, an NSF funded project on bioninformatics (see the BioSIFTER site: http://xtasy.slis.indiana.edu/biosifter/) that involves computer scientists, biologists, and information scientists from both IUPUI and IUB campuses will utilize this infrastructure.

The infrastructure under development differs from existing resources by: (1) its uniform, modular, open architecture; (2) its scalability to handle GB size data sets; (3) its parallel computing infrastructure; (4) its usage of XML-based, OAI derived communication protocols for easy integration of new databases and services as well as the serialization of software packages; (5) detailed documentation of data and code but also links to related publications; and (6) its online GUI supporting the request and navigation of diverse information processing jobs for teaching and research purposes.

References

1. Battista, G., Eades, P., Tamassia, R. and Tollis, I.G. Algorithms for drawing graphs: An annotated bibliography. *Computational Geometry: Theory and Applications*, *4* (5). 235-282.
2. Baumgartner, J. and Börner, K., Towards an XML Toolkit for a Software Repository Supporting Information Visualization Education. in *IEEE Information Visualization Conference*, (Boston, MA, 2002).
3. Bederson, B., Meyer, J. and Good, L., Jazz: An Extensible Zoomable User Interface Graphics Toolkit in Java. in *UIST*, (2000), ACM Press, 171-180.
4. Börner, K. and Zhou, Y., A Software Repository for Education and Research in Information Visualization. in *Information Visualisation Conference*, (London, England, 2001), IEEE Press, 257-262.
5. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41* (6). 391-407.
6. Furnas, G.W., Generalized fisheye views. in *CHI '86*, (1986), ACM Press, 16-23.
7. Kohonen, T., The self-organizing map. in *Proc. IEEE, vol. 73*, (1985), 1551-1558.
8. Kruskal, J.B. Multidimensional scaling and other methods for discovering structure. in Wilf, H. ed. *Statistical Methods for Digital Computers*, Wiley, New York, 1977.
9. Lamping, J., Rao, R. and Pirolli, P., A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. in *ACM CHI'95 Conference on Human Factors in Computing Systems*, (1995), 401-408.
10. Landauer, T.K., Foltz, P.W. and Laham, D. Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*. 259-284.
11. Salton, G., Yang, C. and Wong, A. A vector space model for automatic indexing. *Communications of the ACM*, *18* (11). 613-620.
12. Sarkar, M. and Brown, M.H. Graphical fisheye views. *Communications of the ACM*, *37* (12). 73-84.
13. Schvaneveldt, R.W. *Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood*. Ablex Publishing, NJ, 1990.
14. Shepard, R.N. Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*. 390-398.
15. Shneiderman, B. Tree visualization with tree-maps: A 2-d space filling approach. *ACM Transactions on Graphics*, *11* (1). 92 - 99.
16. Shneiderman, B., Feldman, D., Rose, A. and Grau, X.F., Visualizing digital library search results with categorical and hierarchical axes. in *Fifth ACM conference on ACM Digital Libraries*, (San Antonio, TX, 2000), 57-66.