InfoVis Cyberinfrastructure

Katy Börner
School of Library and Information Science

INDIANA UNIVERSITY
BLOOMINGTON

katy@indiana.edu

SLIS Colloquium, November 19th, 2004

http://iv.slis.indiana.edu/db
http://iv.slis.indiana.edu/cr
http://iv.slis.indiana.edu/sw
http://iv.slis.indiana.edu/lm
Motivation

IVC Database
- Provide access to major scholarly databases.

IVC Software Framework
- Support developers and programmers in the comparison and distribution of new algorithms.
- Interconnect algorithm developers and users. What algorithms do users need/want?

IVC Learning Modules
- Support (non-programmer) users in the utilization of advanced InfoVis algorithms.
- Provide a unique resource for InfoVis education.

Support InfoVis & ‘Knowledge Domain Visualization’ research.

Publications about the Infrastructure

- Penumarthy, Shashikant, Börner, Katy and Herr, Bruce. Information Visualization Cyberinfrastructure Software Framework. Submitted to Information Visualization.

Moral:
Do not do infrastructure development if you need/want scholarly publications.
Grants

- Center of Excellence for Computational Diagnostics. 21st Century Grant (Susanne Ragg, David Clemmer, Sven Rahmann, and Ilka Ott, Terry Vik, R Clement McDonald, Nunroe Pecock, Zina Ben Miled & Katy Börner, $1,994,951) Sept. 04 - Aug. 06.
- Outstanding Junior Faculty Award. (Principal Investigator, $14,000), 2004.
- Sun Center of Excellence in Knowledge Management and Discovery, SUN Microsystems (Stephanie Burks, Katy Börner, Zina Ben-Miled), March 2004.
- Information Visualization Learning Modules. SBC (formerly Ameritech) Fellow Grant (Principal Investigator, $15,000) May 2003-June 2004.

Moral:
A good infrastructure (development) attracts grant funding.
IVC Database – The Team

Design and Implementation

Jay Askren
Saiful Bahari
Andrew Bangert
Christopher Friend
Stephanie Gato

Todd Holloway (Lead)
Ruchi Kapoor
Ketan Mane
Lalitha Visvanath
Qian Wang

Jose Montalvo
Elijah Wright

Graphic Design
Caroline Courtney

Project Start
September 2003

IVC Database - System Overview

Oracle/Apache/Tomcat/Java—Well understood and reliable tools

Oracle DB
- Several terabytes of data
- Relational design
- Allow for more collections to be added

Search Engine
- Search on abstract, author, title, journal, date published, and more
- User login for both IU and non-IU users
- User histories
- Administration of data and user accounts
- Compressed downloading of results
- Term-by-document and co-author matrices of results
Katy Börner: InfoVis Cyberinfrastructure, November 19th, 2004

IVC Database - Data Sets
(http://iv.slis.indiana.edu/db)

Papers and Patents

**NLM**
Number of Entries: 11,601,477
Years Covered: 1953-2001
Size: 135 MB (unzipable)

**PNAS**
Number of Entries: 16,160
Years Covered: 1907-2002
Size: 503 MB

**United States Patent and Trademark Office (Patents)**
Number of Entries: 2,572,447
Years Covered: 1976-2001
Size: 200 MB

Grant Awards

**National Science Foundation (NSF)**
Number of Entries: 191,122
Years Covered: 1965-2002
Size: 400 MB

**National Institute of Health (NIH)**
Number of Entries: 1,553,520
Size: 23 MB

Funding Opportunities

**Community of Science (COS)**
Number of Entries: 36,156 (2,000 new entries per month)
Years Covered: 1990-present
Size: 60 MB
Master Minds/Programmers
Jason Baumgartner, SLIS
Nathan James Deckard, CS
Nihar Sheth, Informatics
Bruce William Herr, CS
Shashikant Penumarthi, SLIS

Algorithm Development and Integration
Vivek Agrawal, Summer Intern
Renee LeBeau, SLIS
Josh Bonner, CS
Todd Holloway, CS
Jeegar Maru, CS
Laura Northrup, CS
Sriram Raghuraman, Informatics
Nihar Sanghvi, Informatics
Hardik Sheth, Informatics
Sidharth Thakur, CS
Ning Yu, SLIS
Yuezheng Zhou, CS

Graphic Design
Caroline Courtney, Fine Art

Students taking K. R. Subramanian’s (UNC Charlotte) InfoVis class integrated diverse algorithms into the IVC.

Project Start
2001

Software
- XML Toolkit
- Preprocessing
- Data Mining
- Layout Algorithms
- Social Visualizations
- Interaction Algorithms
- Other Resources

This page provides pointers to commonly used data analysis and visualization algorithms. An XML Toolkit was implemented to facilitate the efficient visualization of diverse data sets as well as an easy comparison of visualizations generated by different algorithms. The toolkit provides a unified architecture in which algorithms can be easily incorporated. Many software packages are available in Java and hence can be run on any platform that supports Java 1.4.

Most software packages come with:
- Algorithm Description
- Pros & Cons
- Sample Applications
- Implementation Details
- Usage Hints
- References
- Acknowledgements
Preprocessing

- Parsers & Converters
- Stop Word Removal
- Porter Stemming Algorithm
- NICE stemmer

Data Mining

**Vector Space Model**
Developed by Gerard Salton
Soon to be in the XML Toolkit

**Latent Semantic Analysis**
Developed by Tom Griffith and Susan Dumais
Code in XML Toolkit
Original code by Michael Berry

**Topic Model**
Developed by Tom Griffith & Mark Steyvers
Soon to be available via the XML Toolkit

**Burst Detection**
Developed by Jan Hering
Soon to be available via XML Toolkit

**Pathfinder Network Scaling**
Developed by Roger Schmeink
KNOX Tools for Pathfinder Network Analysis can be found via treedepth inc.
Multidimensional Scaling
Developed by Roger N. Shepard
Fast nonlinear MDS algorithm by Matthew Chase and Akbar Memisevic will soon be available in the InfoVis Toolkit.

Self-Organizing Maps
Developed by Tomas Kohonen.
Original code from the VESSEON research group.

Clustering: Ward’s Algorithm
Developed by Ward.
Code in InfoVis Toolkit.

Clustering: Betweenness Centrality
Developed by Dr. Bader.
Code in InfoVis Toolkit.

Layout Algorithms

Parallel Coordinates
Developed by A. Inselberg.
Soon to be available via InfoVis Toolkit.

SimVis
Developed by Yuezhong Zhou.
Code in InfoVis Toolkit.

Spring Embedding Algorithm
Developed by Eades.
Code in InfoVis Toolkit.
Software Packages

GeoZui3D
Developed by Cao, Wen and Michael O'Sullivan.
It is a ZitGraph based visualization tool. It is available for non-commercial use.

Workshop & User Tracked Chat Log Visualizations
Developed by Shahrzad Fazyli and Katy Börner at the Infra Lab at IU.
It is available for use and documented online.

Hierarchical Cluster Explorer
Developed by Jianmin Pan and Michael O'Sullivan.
Original code available at: http://hclab.indiana.edu/

Time Searcher
Developed by Henry Mohler and Ben B. Shneiderman.
Original code available at: http://hclab.indiana.edu/

Open Source Toolkits

IBM's InfoVis Toolkit
by Jean-Francois Fabre.
Interactive Graphics Toolkit written in Java to ease the development of information visualization applications and components.

University of Maryland's Piccolo Toolkit
by Jesse Stephen and Ben Shneiderman at the HCLab.
Piccolo is a toolkit for the creation of robust graphical applications with features such as zooming and multiple representation.
It is based on the Javelin and available as OSI Certified Open Source Software.

AT&T's GraphViz
Set of graph drawing tools for Unix and Windows (Web/MS), including a web-servie interface (webdav).
Source code and binary executables for common platforms are available.

UC Irvine's Java Universal Network/Graph Framework (JUNG)
by David Blaisdell, Joshua O'Mahony, David Fagar, and Jean-Francois Fabre.
Java-based open-source software library designed to support the modeling, analysis, and visualization of data that can be represented as graphs.
It comprises a wealth of algorithms developed in the fields of social network analysis, information visualization, knowledge discovery and data mining.

UC Berkeley's User Interface Research Group's eBeam
by Japheth and Alan Newburger.
Advised by Stuart K. Card and James A. Landay.
Leaf interface tool to building highly interactive visualizations of structured and unstructured data.
Downloads via Sourceforge since June 21, 2004
(http://sourceforge.net/projects/ivc)

Statistics for the past 6 months.

<table>
<thead>
<tr>
<th>Month</th>
<th>Rank (Page Views)</th>
<th>D/I</th>
</tr>
</thead>
<tbody>
<tr>
<td>November 2004</td>
<td>4826 (71.73)</td>
<td>573</td>
</tr>
<tr>
<td>October 2004</td>
<td>6590 (62.05)</td>
<td>501</td>
</tr>
<tr>
<td>September 2004</td>
<td>4900 (70.05)</td>
<td>463</td>
</tr>
<tr>
<td>August 2004</td>
<td>5520 (67.99)</td>
<td>222</td>
</tr>
<tr>
<td>July 2004</td>
<td>4820 (60.00)</td>
<td>236</td>
</tr>
<tr>
<td>June 2004</td>
<td>1081 (44.24)</td>
<td>317</td>
</tr>
</tbody>
</table>

Statistics for All Time

<table>
<thead>
<tr>
<th>Lifespan</th>
<th>Rank</th>
<th>Page Views</th>
<th>D/I</th>
</tr>
</thead>
<tbody>
<tr>
<td>149 days</td>
<td>4826</td>
<td>5,977</td>
<td>573</td>
</tr>
</tbody>
</table>

Workshop on Information Visualization Software Infrastructures

Sat. Oct. 9th, 2004, 10am–4pm
Room 310A, A-B

Right before the IEEE Symposium on Information Visualization (InfoVis) 2004.

Workshop Chairs:
- Jean-Daniel Fekete, INRIA Palaiseau, France (Author of The Infra Toolkit).
  [Jean-Daniel.Fekete@inria.fr](mailto:Jean-Daniel.Fekete@inria.fr)
- Katy Börner, Indiana University, USA (Co-Author of the Infra Toolkit).
  [Katy@indiana.edu](mailto:Katy@indiana.edu)

Description

Information visualization systems and toolkits are becoming available for a large range of visualization and interaction techniques and are used in diverse application domains. This workshop is aimed at getting experts involved in building such infrastructures to share their views, understand the issues involved and try to find ways to avoid fragmentation and improve collaborations.

To participate in the workshop, you should submit a semi-structured position paper explaining your view of what an infrastructure should provide, describe what you consider as the main challenges for such infrastructures and describe the capabilities of toolkits of systems you have already built, following a form available [here](http://vw.indiana.edu/ivsi2004/) by Sept 30th, 2004.

[http://vw.indiana.edu/ivsi2004/]
Katy Börner: InfoVis CyberInfrastructure, SLIS Colloquium, November 19th, 2004

The Information Visualization CyberInfrastructure (IVC) provides access to data, software tools, and learning modules for computer scientists and related researchers to support the analysis, modeling, and visualization of diverse data sets. The IVC Learning Modules (http://iv.slis.indiana.edu/iv) offer several web-based modules that can be used to explore and understand the data. Each module includes an introduction, discussion of existing algorithms, learning tasks, programming exercises, opportunities & challenges, references to research papers and software, and a list of contact information. The software tools and learning modules are designed to support research and education in information visualization.
Visualizing Tree Data

http://iv.slis.indiana.edu/lm/lm-trees.html

Student’s Project Results

User & Task Analysis for Visualizing Tree Data
- Visualizing the structure of IU’s Decision Support System
- Visualizing the co-occurrences of keywords in DLib Magazine articles.
- Visualization of the Java API
- Visualizing the the Library of Congress Classification System to retrieve legal materials in a library.

See Handin pages at
http://ella.slis.indiana.edu/~katy/handin/1579-S04/cgi/handinlogin.cgi

Image by Peter Hook and Rongke Gao
Time Series Analysis & Visualization

http://iv.slis.indiana.edu/lm/lm-time-series.html

Student's Project Results

Time Series Analysis & Visualization

- Using Timesearcher and the Burst Detection Algorithm to Analyze the Stock Market from 1925 to 1945
- Applying Burst and TimeSearcher to Chat Data
- Lab Access Trends
- Quest Atlantis Chat Log Data

See Handin pages at
http://ella.slis.indiana.edu/~katy/handin/L579.S04/cgi/handinlogin.cgi
Visualizing the Work of the United States Supreme Court Based on Time Data and Top Level West Topics

by Peter A. Hook & Rongke Gao

Visualizing the Work of the United States Supreme Court Based on Time Data and Top Level West Topics

by Peter A. Hook & Rongke Gao

Top fifteen most occurring topics from 1944 to 2004 in Timesearcher

All topics grouped by West Category and Sub-Category grouped over the entire lengths of the data set

All topics grouped by West Category and Sub-Category grouped corresponding to the five chief justices

Visualizing Niches of the Blog Universe

BY Mike Tyworth and Elijah Wright

Visualizing Niches of the Blog Universe

BY Mike Tyworth and Elijah Wright

All topics grouped by West Category and Sub-Category grouped over the entire lengths of the data set

All topics grouped by West Category and Sub-Category grouped corresponding to the five chief justices
This course covers

- Perceptual basis of information visualization.
- Data mining algorithms that enable extraction of relationships in data.
- Visualization and interaction techniques.
- Discussions of systems that drive research and development, and
- Future trends and remaining fundamental problems in the field.

Students do weekly readings, provide a presentation on specific readings, do projects, and participate in class & online discussion.

Class Webpage: [http://ella.slis.indiana.edu/~katy/L579](http://ella.slis.indiana.edu/~katy/L579)

---

This course introduces students to major methods, theories, and applications of structural data mining and modeling.

- Covers elementary graph theory and matrix algebra, data collection, structural data mining, data modeling, and applications.

Upon taking this course students will be able to analyze and describe real networks (power grids, WWW, social networks, etc.) as well as relevant phenomena such as disease propagation, search, organizational performance, social power, and the diffusion of innovations.

Format: Lectures and 4-5 labs.
Class Webpage: [http://ella.slis.indiana.edu/~katy/L597](http://ella.slis.indiana.edu/~katy/L597)
Future Work

IVC Database
- Create tables/upload Citeseer, 110 year Physical Review journals dataset, etc.
- Optimize online interface and make it available to other researchers.
- Create connections to R and other packages for large scale (network) data analysis.

IVC Software Framework
- Release IVC core as ‘alpha’.
- Integrate a lot more algorithms.

IVC Learning Modules
- Write new learning modules as new algorithms become available.
- User test learning modules.

Outreach
There will be a Data Analysis, Modeling and Visualization Tutorial @ Electronic Imaging, San Jose, CA, Jan 16th, 2005 which uses the IVC infrastructure.

Do RESEARCH using this infrastructure!

Acknowledgements

The Information Visualization Software Repository was created in 2002 and has since then been used to teach the Information Visualization class at Indiana University. Katy Börner, Yuezheng Zhou, and Jason Baumgartner implemented the very first algorithms. In Summer 2003, Jason Baumgartner, Nilbar Gök, and Nathan J. Deckard lead a project to design a JML tool kit that enables the serialization and parallelization of commonly used data analysis and visualization algorithms. Contributions of software packages and implementation work are acknowledged on the respective software pages. Support comes from the School of Library and Information Science, Indiana University's High Performance Network Applications Program, an Academic Equipment Grant by SUN Microsystems, SRC (formerly Amelco) Fellow Grant, and the National Science Foundation under DUE-0335823 and IIS-0335861.

Craig A. Stewart, Mary Papakhian, Anurag Shankar all UITS generously made the Research Database Complex available for this project and provided very insightful comments.

Stephanie Burks, Principal Unix Systems Administrator, Research and Technical Services, UITS has been instrumental in setting up the computing infrastructure and administration of the Oracle database.