Large Scale Information Visualization

Jing Yang
Fall 2007

Social Visualization

Reference: A large number of slides in this class come from John Stasko’s Infovis class slides. They are used with his permission.
Definition

- Social Visualization
  - “Visualization of social information for social purposes”
    ---Judith Donath, MIT
  - Visualizing data that concerns people or is somehow people-centered

Example Domains

- Social visualization might depict
  - Baby names
  - Conversations
  - Newsgroup activities
  - Email patterns
  - Chat room activities
  - Presence at specific locations
  - Social networks
  - Life histories
Baby Name Visualization

- Baby Names, Visualization, and Social Data Analysis [Wattenberg Infovis 2005]
- NameVoyager – a web-based visualization applet
  - Let users interactively explore name data, historical name popularity figures
  - [http://babynamewizard.com/namevoyager/lnv0105.html](http://babynamewizard.com/namevoyager/lnv0105.html)
  - More than 500,000 site visits in the first two weeks
  - Average of 10,000 visits per day after two months
- Lesson – To design a successful exploratory data analysis tool, one good strategy is to create a system that enables “social” data analysis

Baby Name Visualization

- Similar to Themeriver
Social Network Visualization

- Vizster: Visualizing Online Social Networks [Heer Infovis 05]
- Online social networks – millions of members publicly articulate mutual “friendship” relations
  - Friendser.com, Tribe.net, and orkut.com
- Vizster
  - Playful end-user exploration and navigation of large-scale online social networks
  - Explore connectivity, support visual search and analysis, and automatically identifying and visualizing community structures
  - Video

Social Network Visualization

- Vizster: Visualizing Online Social Networks [Heer Infovis 05]
- Usage observation
  - 500-person all-night event
  - Many party-goers are familiar with the friendster system
  - Interactive kiosk and a projection of the visualization onto a large screen
Email Visualization

- THREAD ARCS: An Email Thread Visualization [Kerr Infovis 2003]
- Thread Arcs
- Combine the chronology of messages with the structure of a conversational thread
- Help people learn various attributes of conversations and find relevant messages

THREAD ARCS: An Email Thread Visualization [Kerr Infovis 2003]

- Basic ideas:

  Figure 1. Chronology of message nodes in a line of six messages.

  Figure 2. Relationships are shown with "reply to" arcs connecting nodes above the line.
THREAD ARCS: An Email Thread Visualization [Kerr Infovis 2003]

Design choices:

Figure 4. The relationship between messages are clearer when arcs are drawn above and below nodes (B).

Figure 7. Unconstrained (A) and constrained arc heights (B) for a thread of sixteen messages. A selection highlighting scheme for the ninth message (C).

Highlight strategies:

Figure 16. Different message attributes highlighting schemes.
THREAD ARCS: An Email Thread Visualization [Kerr Infovis 2003]

Prototype:

![Prototype Image]

Figure 18. Thread Arcs integrated into email client prototype.

Chat Room Visualization

Chat Circles [Viegas and Donath CHI'99]

http://chatcircles.media.mit.edu/about.html

You can try it out!

- GUI for chat rooms
- Represent people using circles
- Mimics cocktail party in certain ways
Chat Circles [Viegas and Donath CHI’99]
Chat Circles [Viegas and Donath CHI’99]

- Each participant is a colored circle
- Circle grows with each posted message, slowly shrinks/fades as goes idle
- Will stay there as small circle while connected
- Comments appear inside circles
- Can only “hear” what is going on nearby

Chat Circles [Viegas and Donath CHI’99]

- History interface
Chat Circles [Viegas and Donath CHI’99]

- Mapping
  - Individual users on x-axis
  - Time goes up on y-axis
  - Tick marks are postings, mouse over reveals them
  - Solid tick marks were within earshot of you, hollow ones weren’t

- Try it live
  http://chatcircles.media.mit.edu/

Chat Circles [Viegas and Donath CHI’99]

- Each participant is a colored circle
- Circle grows with each posted message, slowly shrinks/fades as goes idle
- Will stay there as small circle while connected
- Comments appear inside circles
- Can only “hear” what is going on nearby
Discussion Group Visualization

- Discussion group:
  - Web-based message boards
  - Usenet newsgroups
  - Chatrooms

- Questions:
  - Do participants really get involved?
  - How much interaction is there?
  - Do participants welcome newcomers?
  - Who are the experts?

People Garden [Xiong and Donath UIST’99]

- Visualization technique for portraying online interaction environments (Virtual Communities)
- Provides both individual and societal views
- Utilizes garden and flower metaphors
Data Portrait: Petals

- Fundamental view of an individual

  ![Petals Diagram]

  His/Her postings are represented as petals of the flower, arranged by time in a clockwise

Data Portrait: Postings

- Time of Posting

  ![Postings Diagram]

  - New posts are added to the right
  - Slide everything back so it stays symmetric
  - Each petal fades over time showing time since posting
  - A marked difference in saturation of adjacent petals denotes a gap in posting
Data Portrait: Responses

- Small circle drawn on top of a posting to represent each follow-up response

Data Portrait: Color

- Initial post vs. reply
  - Color can represent original/reply
  - Here magenta is original post, blue is reply
Garden

- Combine many portraits to make a garden
- Message board with 1200 postings over 2 months
- Each flower is a different user
- Height indicates length of time at the board

Alternate Garden View

- Sorted by number of postings
Interpreting Displays

Group with one dominating person

More democratic group

Text and Document Visualization

Class 12, Part C
Text is Everywhere

- We use documents as primary information artifact in our lives
- Our access to documents has grown tremendously in recent years due to networking infrastructure
  - WWW
  - Digital libraries
  - ...

Big Question

- What can information visualization provide to help users in gathering information from text and document collections?
InfoVis Tasks

- Two main tasks that Information Visualization can assist with in this area
  - Enhance a person's ability to read, understand and gain knowledge from a document
  - Understand the contents of a document or collection of documents without reading them

Specific Tasks for Document Collections

- What are the main themes of a document?
- How are certain words or themes distributed through a document?
- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that might be close enough to be worthwhile?
### Simple Taxonomy

<table>
<thead>
<tr>
<th></th>
<th>Enhanced presentation (syntax)</th>
<th>Concepts and relationships (semantics)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single document</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collection of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>documents</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Enhanced Presentation of a Document

Text is too small to read
Enhanced Presentation of a Document

Magnifying lens

Fisheye view

Bifocal display

Perspective wall

Enhanced Presentation of a Document

Document Lens
Enhanced Presentation of a Document

Document Lens

Enhanced Presentation of a Document

Zoom Browser
Enhanced Presentation of Labels

- Dynamic Visualization of Graphs with Extended Labels [Wong et al. Infovis 2005]

video

Enhanced Presentation of Labels

- Excentric Labeling [Fekete and Plaisant CHI '99]
Concepts and Relationships in Individual Document

- TOPIC ISLANDSTM – A Wavelet-Based Text Visualization System [Miller Vis’ 98]
  - Construct digital signals from words within a document
  - Apply wavelet transforms to the signals
  - Analyze narrative flow using resultant wavelet energy
  - Use MDS to map themes

Topic Islands [Miller Vis 98]

- Construct digital signals from words within a document
  - Channels or topics: content-bearing
  - Signal for the channels are stored.
  - Wavelet transforms are applied to the signals to calculate three types of wavelet energy:
    - Channel energy: signal for each channel is processed independently
    - Composite energy: include all information across all channels
    - Query energy: show local relevance between narrative and query.
Topic Islands [Miller Vis 98]

- Composite energy: high frequency – break point

![Figure 2: Detail strip chart for Buddhist article. Final break points are shown as white ticks.](image)

Topic Islands [Miller Vis 98]

- subchunk position: MDS of themes
- subchunk base sizes: length or other variables

![Figure 3: Island view for Buddhist article created from document sub-chunking shown in Figure 2.](image)
Figure 14: The top view of the islands for the Buddhist Article shows the interplay of theme map. See also CP 8.

Figure 13: The top view of the islands for the Ireland article show the interplay of theme map and the narrative. See CP7.
Document Collections

- Problem or challenge is how to present the contents/semantics/themes/etc of the documents to someone who does not have time to read them all
- Who cares?
  - Researchers, news people,…

Improving Text Searches

- What’s wrong with the common search?
  - Query responses do not include:
    - How strong the match is
    - How frequent each term is
    - How each term is distributed in the document
    - Overlap between terms
    - Length of document
  - Document ranking is opaque
  - Inability to compare between results
  - Input limits term relationships
TileBars [Hearst CHI’95]

- **Goal**
  - Minimize time and effort for deciding which documents to view in detail

- **Idea**
  - Show the role of the query terms in the retrieved documents, making use of document structure

---

TileBars [Hearst CHI’95]

- **Techniques**

  - Relative length of document
  - Two search terms
  - Blocks indicate “chunks” of text, such as paragraphs
  - Blocks are darkened according to the frequency of the term in the document
TileBars [Hearst CHI’95]

http://elib.cs.berkeley.edu/tilebars/about.html#using

More Complex Process
Visualizing Documents

- Break each document into its words
- Two documents are “similar” if they share many words
- Use algorithm for clustering similar documents together and dissimilar documents far apart

Use SOM Map
IN-SPIRE

- Document visualization and analyzing system by PNNL
- Enable users to review and analyze thousands of documents simultaneously using interactive, visually oriented framework
- Requires almost no advanced knowledge of the information that is being processed
- Provide overview: "Lay-of-the-land" from a topical perspective.
- Provide query and display tools to support deeper analysis and interrogation of the information space.

Galaxy Overview
ThemeScape Overview

Interactive Exploration

- Automatic data foraging
- Document analysis: diagnose, outlier term removal, correlation analysis, full text
- Dynamic Layout: re-MDS for subgroups and updated term sets
- Time related analysis
- Document search: query by keywords, query by example, group query result
- Query organization: save/load query, query history
- Document organization: add group, highlight group, group from query result
- Evidence organization and exchange
WebTheme

ThemeRiver
Image and Video Visualization

Goals
- 1. Enhance visualization of individual images
- 2. Using visualization to improve automated analysis of image and video collections
- 3. Using visualization to help users browse and search image and video collections

DragMag
- Enhance visualization of high resolution images
- video
Image and Video Collections

- Assume you have a big image or video collection:
  1. What kind of tasks can be related to the collection?
  2. What kind of information about the photos can be used?
  3. What kind of information visualization techniques can be used to perform tasks?

Available Information

- Manual annotation, ranks
- Hierarchical structure of the collection, folders
- Semi-automatic annotation
  - Geographic Positioning Systems (GPS) and date recorders built in cameras.
- Detected patterns from Automatic analysis
  - Visual attributes
  - Semantic annotations
Examples

- Video: Photo Finder  - Ben Shneiderman et al.
  - Multiple views of organized photo collections
- Video: Phote Mesa    - Benjamin Bederson
  - Zoomable image browser
- Video: Timequilt    - Patrick Baudisch et al.
  - Zoomable image browser with temporal information conveyed by image layout

Examples (con.)

- Video: Semantic image browser - Jing Yang et al.
  - Interactive exploration using semantic information
- Video: Image analysis application: Satellite images  - NASA
  - Convert to high dimensional dataset
- Video: LensRiver
  - news video browsing
WWW and Internet Visualization

Reference: John Stasko’s Infovis class; http://www.cs.brown.edu/memex/ACMCSHT/51/51.html

Motivation

- Aid authors and webmasters with production and organization of content
- Assist Web surfers making sense of the information
- Help researchers understand the Web
Main Topics

- Presentations of the Internet and WWW
  - Focus on topology and navigation, similar to the graph visualization work
  - Visualizing the evolution of the Web
- Visual aids for browsing and using the WWW and the Internet
  - Assistive visualizations not focusing on presenting net structure and connectivity
  - Visualizing clickstreams
  - Visualizing users
  - Visualizing searches

Major Challenges

- Websites simply are too big
- Huge graphs
- Layout is challenging
CAIDA

- CAIDA: Cooperative Association for Internet Data Analysis
  - provides tools and analyses promoting the engineering and maintenance of a robust, scalable global Internet infrastructure
  - [http://www.caida.org/home/](http://www.caida.org/home/)
CAIDA Macroscopic Topology Measurements project

- Goal: measures connectivity and latency data for a wide cross-section of the commodity Internet
- Method: track global IP level connectivity by sending probe packets from a set of source monitors to hundreds of thousands of destinations stratifying the current IPv4 address space as well as the Earth
- Visualizations are used to analyze collected data, especially network connectivity

http://www.caida.org/analysis/topology/macroscopic/
http://www.caida.org/tools/measurement/skitter/visualizations.xml

AS Internet graph

- Input: IP addresses and IP links
  - topology of Autonomous Systems (ASes). Each IP address are mapped to the AS responsible for routing it
- Layout: the position of each AS node is plotted in polar coordinates

\[
\text{radius} = 1 - \log \left( \frac{\text{outdegree}(\text{AS}) + 1}{\text{maximum.outdegree} + 1} \right)
\]

\[
\theta = \left( \frac{\text{longitude of the AS headquarters}}{\text{in whois records}} \right)
\]

- Outdegree: number of "next hop" ASes observed accepting traffic from this AS
AS Internet graph

The link color reflects outdegree, from lowest (blue) to highest (yellow).

Insight:
• ISPs in Europe and Asia have many peering relationships with ISPs in the U.S. there are fewer links directly between ISPs in Asia and Europe.

AS Paths in Hypviewer
Visualizing the Global Topology of the MBone

- MBone: the Internet's multicast backbone
  - Multicast: distributing data from one sender to multiple receivers with minimal packet duplication
- MBone has been extremely popular for efficient transmission across the Internet of real-time video and audio streams such as conferences, meetings, congressional sessions, and NASA shuttle launches
- MBone grew exponentially with no central authority
- video
- Tamara Munzner and Eric Hoffman and K. Claffy and Bill Fenner

Natto [Shiozawa and Matsushita HCI International '97]

- Target: web pages with links
- Initial layout: a flat horizontal plane.
- Node placement: map attributes of the web page (e.g. its size, title, number of images) to the two-axis of the plane.
- Interaction: users can select nodes and raise them vertically to de-occlude the structure. Adjacent (linked) nodes maintain a close proximity to the raised nodes so that the structure is gradually "disentangled" from the plane
- Limitations: the number of nodes that may comfortably occupy the flat plane before selection
Natto [Shiozawa and Matsushita HCI International '97]

Open Text Web Index [Tim Bray 1996]
- Visualize websites as composite objects placed in 3D space
- Represent variables associated with web sites, such as number of pages, number of links to and from these pages, and domain identifiers
- Use the distance between two objects to represent the degree of connectivity between the two sites
- Work as a map in the site level
Open Text Web Index [Tim Bray 1996]

MAPA [Durand and Kahn Proc. ACM Hypertext '98]
- Aims to improve navigation in large web sites of between 500 and 50,000 pages
- Presents pages as square icons that stand in rows and columns on a flat plane
- A focus page is placed at the front edge of the plane and its child pages form a row behind the focus page. Each child page then has its children behind it so as to form a single column behind the page
MAPA [Durand and Kahn Proc. ACM Hypertext '98]

Visual aids for browsing and using the WWW and the Internet

- Potential web-related tasks
  - How and when has info been accessed?
  - Where do people enter and spend time?
  - How do they move about?
  - What paths aren't traversed?
  - Where are they coming from?
  - What has been added, changed, deleted?
  - Do changes affect navigation patterns?
  - Do we need to do a redesign?
Data Set

- Each server request is a data case
- Example variables
  - IP Address/Client host
  - Timestamp
  - URL requested
  - HTTP status (success, not found, …)
  - Bytes delivered
  - Referencing URL (HTTP-Referrer)
  - User agent (browser and OS info)

E-Commerce Clickstream Visualization [Brainerd and Becker Infovis 2001]

- Brainerd and Becker Infovis 2001
- Goal: analyzing user behavior of a web site
  - Understand the interactions between users and web site
- Visualization:
  - shows site topology and traffic flow
  - Presents a more complete picture of web site usage by segmenting site traffic data based on user attributes, including demographic data and purchase history
Gender Differences: Males tend to navigate in specific, direct patterns, whereas women's navigation patterns include more browsing, utilizing more of the site.

Checkout process. Purchasers take a direct route through the checkout process, whereas non-purchasers show a more haphazard route, including self-edges and early abandonment, possibly indicating a confusing checkout process.
Biologists use high-throughput experiments to answer complex biological research questions. Experiments, such as gene-expression microarrays, result in datasets that are very large. Reference: An evaluation comparing microarray tools [Saraiya et al InfoVis 2004]
Microarray Dataset Examples

- Time Series - Data for 1060 genes over 5 time points of a viral infection cycle in human embryonic kidney cells (1060 rows, 5 columns)
- Viral Conditions - Data for 861 genes for 3 related viral infections at 8 hrs post infection of human lung epithelial cells (861 rows, 3 columns)
- Lupus vs. Control - Data for 170 genes from 42 control (healthy) people and 48 people suffering from systemic lupus erythematosus (SLE), an autoimmune disease (170 rows, 90 columns)

Clusterview

- Heat-map
  - Increased gene-expression values: red brightness scale
  - Decreased gene-expression values: green brightness scale
  - No change: black.
TimeSearcher

Parallel Coordinates for both overview and individual views
Dynamic query for filtering

Hierarchical Clustering Explorer

Dendrogram visualization with heat-map
Spotfire

Place each cluster in a separated parallel coordinates window

GeneSpring

The largest variety of visualizations for microarray data analysis
Definition

“The use of the crafts of typography, graphic design, animation, and cinematography with modern human computer interaction and computer graphics technology to facilitate both the human understanding and effective use of computer software.”

Price, Baecker and Small, ’98
Challenge

- Software clearly is abstract data
- Unlike much information visualization, however, software is often dynamic, thus requiring our visualizations reflect the time dimension
  - History views
  - Animation
  - ...

Sub-domains

- Two main sub-areas of software visualization
  - Program visualization - Use of visualization to help programmers, coders, developers. Software engineering focus
  - Algorithm visualization - Use of visualization to help teach algorithms and data structures. Pedagogy focus
Program Visualization

- Can be as simple as enhanced views of program source
- Can be as complex as views of the execution of a highly parallel program, its data structures, run-time heap, etc.

Enhanced Code Views

```cpp
/**
 * Gets the normalized value of a data.
 * @param valueString the data in a string
 */
double GetNormalizedValue(string valueString) {
    double value = stod(valueString.c_str());
    if (min == max)
        return 0;
    else if (value == min)
        return 0;
    else if (value == max)
        return 1;
    double v = (value - min) / (max - min);
    if (v < 0)
        v = 0;
    else if (v > 1)
        v = 1;
    return v;
```
SeeSoft System [Eick et al. IEEE ToSE ’92]

- Pulled-back, far away view of source code
- Map one line of source to one line of pixels
  - Can indicate line indentation, etc.
  - Use color to represent the programmer, age, or functionality of each line.
- Like taping your source code to the wall, walking far away, then looking back at it

SeeSoft System View
Use

- Tracking (typically means mapping this data attribute to color)
- Code modification (when, by whom)
- Bug fixes
- Code coverage or hotspots
- Interactive, can change color mappings, can brush views, can compare files, …

Tarantula [Eagan et al. Infovis’01]

- Utilizes SeeSoft code view methodology
- Takes results of test suite run and helps developer find program faults
- Clever color mapping is the key!
Color Mapping of Tarantula

- Color reflects a statement’s relative success rate of its execution by the test suite.
  - Color spectrum: from red to yellow to green
- Statements executed by a failed test case become more red
- Statements executed by a passed test case become more green
- Statements shown as red are highly suspect
- Statements shown as green convey a strong confidence in their correctness
- Statements shown as yellow convey a sense of ambiguousness,

Tarantula View

![Tarantula View Image]
Software Structure Visualization

- Call graph visualization
- Flow chart visualization
- Graph visualization!

Sample Call Graph View
FIELD [Reiss Software Pract & Exp’90]

- Program development and analysis environment with a wide assortment of different program views
  - Integrated a variety of UNIX tools
  - Utilized central message server architecture in which tools communicated through message passing

FIELD [Reiss Software Pract & Exp’90]

- Interface
FIELD [Reiss Software Pract & Exp’90]

- Dynamic Call Graph View

- Class browser
FIELD [Reiss Software Pract & Exp’90]

- Heap View
  - Color could be
  - When allocated
  - Block size
  - Where allocated

FIELD [Reiss Software Pract & Exp’90]

- 3D call graph
Multilevel Call Matrices [vanHan Infovis 2003]

Node-link diagram

Multilevel Call Matrices [vanHan Infovis 2003]

Call Matrix

Figure 2: Recursive matrix subdivision according to relative component size (a). Matrix subdivision according to number of subcomponents at a specific abstraction level (b)
Multilevel Call Matrices [vanHan Infovis 2003]

Figure 5: Displaying additional data: call permissions (a), call neighborhood (red calls are closer to call under pointer) and call density (c)

PV System [Kimelman et al. Vis94]

- Used for understanding application and system behavior for purposes of debugging and tuning
- Users look for trends, anomalies, and correlations
- Ran on RISC/6000 workstations using AIX
- Trace-driven, can be viewed on-line or off
Different Views

- Hardware-level performance info
  - Instruction execution rates, cache utilization, processor utilization
- Operating system level activity
  - Context switches, system calls, address space activity
- Communication library level activity
  - Message passing, interprocessor communication
- Language run-time activity
  - Dynamic memory allocation, parallel loop scheduling
- Application-level activity
  - Data structure accesses, algorithm phase transitions
A number of commercial program development environments have begun to incorporate program visualization tools such as these.

- Majority are PC-based
- Has not become wide-spread
Concurrent Programs

- Understanding parallel programs is even more difficult than serial
- Visualization and animation seem naturals for illustrating concurrency
- Temporal mapping of program execution to animation becomes critical

Example system: POLKA [stasko & Kraemer JPDC '93]

Message Passing Systems

- PVM/Conch [Topol et al. JPDSN '98]
Shared Memory Threads

- Pthreads [Zhao & Stasko TR '95]

Algorithm Visualization

- Learning about algorithms is one of the most difficult things for computer science students
  - Very abstract, complex, difficult to grasp
- Idea: Can we make the data and operations of algorithms more concrete to help people understand them?
Algorithm Animation

- Common name for area
- Dynamic visualizations of the operations and data of computer algorithm as it executes

Sorting Out Sorting

- Seminal work in area
- 30 minute video produced by Ron Baecker at Toronto in 1981
- Illustrates and compares nine sorting algorithms as they run on different data sets

- Demo
  
  http://kmdi.utoronto.ca/RMB/publications.html
Binky Pointer Fun Video

- Stanford CS Education Library: Pointer Fun With Binky -- a fun 3 minute video that explains the basics features of pointers and memory

Balsa [M. Brown Computer ’88]

- First main system in area
- Used in “electronic classroom” at Brown
- Introduced use of multiple views and interesting event model
Example Animation

Tango [Stasko Computer ’90]

- Smooth animation
- Simplification of the design/programming Process
- Formal model of the animation
POLKA [Stasko & Kraemer JPDC ’93]

- A general purpose animation system that is particularly well-suited to building algorithm and program animations
- Parallel programs and serial programs
- Provide an interactive, front-end called Samba.
  - Samba is an animation interpreter that reads one ascii command per line, then performs that animation directive. These commands are of the form:
    rectangle 3 0.1 0.9 0.1 0.1 blue solid
    move 3 0.5 0.0

POLKA [Stasko & Kraemer JPDC ’93]

- Improved animation design model
- Object-oriented paradigm
- Multiple animation windows
- Much richer visualization/animation capabilities
A Useful Link

http://www.cc.gatech.edu/gvu/softviz/SoftViz.html