Lecture 09 Interactive Visualization and Visual Analytics

Andreas Holzinger
VO 709.049 Medical Informatics
11.01.2017 11:15-12:45

Science is to test crazy ideas – Engineering is to put these ideas into Business

(a_holzinger@tugraz.at)
http://hci-kdd.org/biomedical-informatics-big-data

00 Reflection – follow-up from last lecture
- 01 Verbal vs. Visual Information
- 02 Informatics as Semiotics Engineering
- 03 Visualization Definitions
- 04 Usefulness of Visualization
- 05 Visualization Methods
(long chapter but incomplete!)
01 apple
Verbal Information

VS.

Visual Information

Problem: Context!
A picture is worth a thousand words?

02 Informatics as Semiotics Engineering

1. Physical: is it present?
   - Signals, traces, components, points, ...
2. Empirical: can it be seen?
   - Patterns, entropy, codes, ...
3. Syntactic: can it be read?
   - Formal structure, logic, deduction, ...
4. Semantic: can it be understood?
   - Meaning, proposition, truth, ...
5. Pragmatic: is it useful?
   - Intentions, negotiations, communications, ...
6. Social: can it be trusted?
   - Beliefs, expectations, culture, ...


03 What is Visualization?

Images are perceived as a set of signs
Sender encodes information in signs
Receiver decodes information from signs
"Resemblance, order and proportion are the 3 "signifieds" in graphics"
"With up to three rows, a data table can be constructed directly as a single image ... However, an image has 3 dimensions And this barrier is impassible."


Visualization is a typical HCI topic!

- the common denominator of Computational sciences
- the transformation of the symbolic into the geometric
- the support of human perception
- facilitating knowledge discovery in data


04 Usefulness of Visualization Science
**05 Visualization Basics**

- **Slide 9-17 A look back into History...**
  - John Snow's map of cholera in London, 1854. Social Science & Medicine, 66, 8, 1266-1269.

- **Slide 9-19 Systematic Visual Analytics > Content Analytics**

- **Example: Data structures - Classification**
  - Data structures:
    - Nominal
    - Ordinal
    - Interval
    - Ratio

- **What do you see in this picture?**
  - Florence Nightingale - first medical quality manager

- **Slide 2-15: Categorization of Data (Classic "scales")**
  - Scale | Empirical Operation | Mathem. Group Structure | Transf. in Il | Basic Statistics | Mathematical Operations
  - --- | --- | --- | --- | --- | ---
  - **Nominal** | Determination of equality | 
  - **Ordinal** | Determination of more/less
  - **Interval** | Determination of equality or differences
  - **Ratio** | Determination of equality or ratios

- **Remember Data structures**
5 Visualization Methods (Incomplete!)
A polygonal line $P$ on the $N-1$ points represents a point $P = (p_1, ..., p_{i-1}, p_i, ..., p_N) \in \ell$

since the pair of values $p_{i-1}, p_i$ marked on the $X_{i-1}$ and $X_i$ axes.

In the following slide we see several polygonal lines, intersecting at $\ell_{(i-1),i}$

representing data points on a line $\ell \subset \mathbb{R}^10$.

Note: The indexing is essential and is important for the visualization of proximity properties such as the minimum distance between a pair of lines.

---

**Why are such approaches not used in enterprise hospital information systems?**
**Tactical Example: Flow Cytometry (3) Immunophenotyping**

- Forward scatter channel (FSC) intensity equates to the particle's size and can also be used to distinguish between cellular debris and living cells.
- Side scatter channel (SSC) provides information about the granular content within a particle.
- Both FSC and SSC are unique for every particle, and a combination of the two may be used to differentiate different cell types in a heterogeneous sample.


**TU Example: Limitations of 2D Parallel Coordinates**

1. Let us consider a point \( y_i = (y_{i1}, y_{i2}, \ldots, y_{id}) \) from the \( d \)-dimensional space.
2. This point is now mapped into a single point \( u \) in the plane of anchors: for each anchor \( j \) the stiffness of its spring is set to \( y_j \).
3. Now the Hooke's law is used to find the point \( u \), where all the springs forces reach equilibrium (means they sum to 0). The position of \( u = (u_1, u_2) \) is now derived by:

\[
\begin{align*}
\sum_{i=1}^{n} \frac{\sum_{j=1}^{d} y_{ij} - u_j}{u_j} y_i &= 0 \\
\sum_{i=1}^{n} \frac{\sum_{j=1}^{d} y_{ij}}{u_j} y_i &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{d} y_{ij} \cos(u_j)
\end{align*}
\]


**TU Parallel Coordinates in 3D**

- **a**
- **b**
- **c**
- **d**

Streit et al. (2006)

**TU Slide 9-32 RadViz – Idea based on Hooke's Law**


**TU Slide 9-35 RadViz for showing the existence of clusters**

Source: http://orange.ويكليب.وي/
Slide 9-37 Star Plot production

- Arrange N axes on a circle in \( \mathbb{R}^2 \)
- \( 3 \leq N \leq N_{\text{max}} \)
  - Note: An amount of \( N_{\text{max}} \leq 20 \) is just useful, according to Lanzenberger et al. (2005)
- Map coordinate vectors \( P \in \mathbb{R}^N \) from \( \mathbb{R}^N \rightarrow \mathbb{R}^2 \)
- \( P = (p_1, p_2, \ldots, p_N) \in \mathbb{R}^N \) where each \( p_i \) represents a different attribute with a different physical unit
- Each axis represents one attribute of data
- Each data record, or data point \( P \) is visualized by a line along the data points
- A line is perceived better than points on the axes

Slide 9-38 Algorithm for drawing the axes and the lines

```javascript
angle = 2 * \pi / N
for each a, from axes:
  angle = a * angle
  x = mid_x + r * cos(angle)
  y = mid_y + r * sin(angle)
  DrawLine(midpoint_x, midpoint_y, x, y)
```

Slide 9-39 Visual Analytics is intelligent HCI

- **Update**
- **Learn**
- **Visual interface**
- **Visual communication**
- **Interact**
- **Visualize**

Slide 9-40 Design of Interactive Information Visualization

1. What facets of the target information should be visualized?
2. What data source should each facet be linked to and what relationship do these facets have?
3. What layout algorithm should be used to visualize each facet?
4. What interaction techniques should be used for each facet and for which functions?

Slide 9-41 Overview first - then zoom and filter on demand

- 1) Overview: Gain an overview about the entire data set (know your data);
- 2) Zoom: Zoom in on items of interest;
- 3) Filter: Filter out uninteresting items — get rid of distractors — eliminate irrelevant information;
- 4) Details-on-demand: Select an item or group and provide details when needed;
- 5) Relate: View relationships among items;
- 6) History: Keep a history of actions to support undo, replay, and progressive refinement;
- 7) Extract: Allow extraction of sub-collections and of the query parameters;

Slide 9-42 Letting the user interactively manipulate the data

- Focus Selection = via direct manipulation and selection tools, e.g., multi-touch (in data space a n-dim location might be indicated);
- Extent Selection = specifying extents for an interaction, e.g., via a vector of values (a range for each data dimension or a set of constraints);
- Interaction type selection = e.g., a pair of menus: one to select the space, and the other to specify the general class of the interaction;
- Interaction level selection = e.g., the magnitude of scaling that will occur at the focal point (via a slider, along with a reset button)

Slide 9-43 Rapid Graphical Summary of Patient Status

- **Date of admission**: today's date
- **Glucose**: 237 mg/dl
- **Blood pressure**: critical elevated elevated normal range reduced critically reduced
- **More than one year prior to admission**: year prior to admission
- **First week of admission**: year prior to admission

Slide 5-44 Example Project Lifelines

What are temporal analysis tasks?

**Classification** - given a set of classes; the aim is to determine which class the dataset belongs to; a classification is often necessary as pre-processing.

**Clustering** - grouping data into clusters based on similarity; the similarity measure is the key aspect of the clustering process.

**Search/Retrieval** - look for a priori specified queries in large datasets (query-by-example); can be exact matched or approximate matched (similarity measures are needed that define the degree of exactness).

**Pattern discovery** - automatically discovering relevant patterns in the data, e.g., local structures in the data or combinations thereof.

**Prediction** - foresees likely future behaviour of data - to infer from the data collected in the past and present how the data will evolve in the future (e.g. autoregressive models, rule-based models etc.)


---

**Remember:**

**Subspace Clustering**

---

**05 Conclusion and Future Outlook**

- **Parsons et al. SIGKDD Explorations 2004**

- **Slide 9-14** We can conclude that Visualization is ...
  - the common denominator of Computational sciences
  - the transformation of the symbolic into the geometric
  - the support of human perception
  - facilitating knowledge discovery in data

- Time (e.g. entropy) and Space (e.g. topology)
- Knowledge Discovery from "unstructured" ;-) (Forrester: >80%) data and applications of structured components as methods to index and organize data -> Content Analytics
- Open data, Big data, sometimes: small data
- Integration in "real-world" (e.g. Hospital), mobile
- How can we measure the benefits of visual analysis as compared to traditional methods?
- Can (and how can) we develop powerful visual analytics tools for the non-expert end user?

Thank you!

Sample Questions (1)
- What is semantic engineering?
- Please explain the process of intelligent interactive information visualization!
- What is the difference between visualization and visual analytics?
- Explain the model of perceptual visual processing according to Ware (2004)
- What was the historical start of systematic visual analytics? Why is this an important example?
- Please describe very shortly 6 of the most important visualization techniques!
- Transform five given data points into parallel coordinates!
- How can you ensure data protection in using parallel coordinates?
- What is the basic idea of RadVis?
- For which problem would you use a star-plot visualization?

Sample Questions (2)
- What are the basic design principles of interactive intelligent visualization?
- What is the visual information seeking mantra of Shneiderman (1996)?
- Which concepts are important to let the end user interactively manipulate the data?
- What is the problem involved in looking at neonatal polysomnographic recordings?
- Why is time very important in medical informatics?
- What was the goal of LifeLines by Plaisant et al (1996)?
- Which temporal analysis tasks can you determine?
- Why is pattern discovery in medical informatics so important?
- What is the aim of foreseeing the future behaviour of medical data?

Some useful links
- http://vis.lbl.gov/Events/SC07/Drosophila/ (some really cool examples of high-dimensional data)
- http://people.cs.uchicago.edu/~wiseman/chernoff (Chernoff Faces in Java)
- http://lib.stat.emu.edu (iris sample data set)
- http://graphics.stanford.edu/data/voldata (113-slice MRI data set of CT studies of cadaver heads)

Appendix: Parallel Coordinates in a Vis Software in R

http://datamining.topware.com
The model of Corchado et al. (2009) combines:
1) methods to reduce the dimensionality of the original data set;
2) pre-processing and data filtering techniques;
3) a clustering method to classify patients; and
4) extraction of knowledge techniques.
The system reflects how human experts work in a lab, but
1) reduces the time for making predictions;
2) reduces the rate of human error; and
3) works with high-dimensional data from exon arrays.