Visual Analytics of Heterogeneous Data in Life Science Applications

Hans-Jörg Schulz
Agenda

I. Motivation, Definitions

II. Visual Analytics of Inhomogeneous Data

III. Orientation and Navigation in Heterogeneous Data

IV. Conclusion, Food for Thought
I. Motivation, Definitions

Linked Data, Open Data

Source: Richard Cyganiak, Anja Jentzsch
I. Motivation, Definitions

Challenges of Multiple Data Sources

- Multiple **origins** (different measurement precisions, different languages,…)
- Multiple **formats** (for querying – SQL, SPARQL,… and for query results – JSON, XML,…)
- Multiple **access restrictions** (and authorities to grant access are also distributed)
- Multiple **data types** (images, documents, numerical values, graphs/structures,…)
- Multiple **data scales** (different value ranges)

- Multiple **analysis methods** (R or Weka for numerical data, Lucene and LingPipe for text,…)
- Multiple **visualization techniques** (image viewers, text visualization, charts+plots,…)

How to do Visual Analytics in such a scenario?
I. Motivation, Definitions

**Heterogeneous Data in Biomedical Applications**

- population
- patient data / clinical data
- MR / CT / X-ray
- tissue samples
- pathway data
- gene expression data
- + pharmaceutical data bases
- + PubMed publications data base
- + disease data bases (ICD-10, DSM-IV, ...)
- + gene and protein data bases (NCBI)
I. Motivation, Definitions

**Heterogeneous Data**

Data drawn from multiple separate data sets/data bases.

**Inhomogeneous Data**

Data from a single data set/data base which is non-uniformly distributed, contains values from different scales, or missing values.
II. Visual Analytics of Inhomogeneous Data

Alexander Lex, Hans-Jörg Schulz, Marc Streit, Christian Partl, and Dieter Schmalstieg: VisBricks: Multiform Visualization of Large, Inhomogeneous Data, appeared at InfoVis’11
II. Visual Analytics of Inhomogeneous Data

Premise

Different (homogeneous) subsets of an inhomogeneous data set

- exhibit different data characteristics
- which must be analyzed differently
- and shown differently

within the context of the whole data set.
II. Visual Analytics of Inhomogeneous Data

Overall Approach: Divide & Conquer

SPLIT-UP DATA → TREAT PARTS INDIVIDUALLY → BRING RESULTS BACK TOGETHER

SPLIT-UP DATA → TREAT PARTS INDIVIDUALLY → BRING RESULTS BACK TOGETHER

SPLIT-UP DATA → TREAT PARTS INDIVIDUALLY → BRING RESULTS BACK TOGETHER
## II. Visual Analytics of Inhomogeneous Data

### The Divide Step: Possible Inhomogeneities in Tabular Data

<table>
<thead>
<tr>
<th><strong>Semantics</strong></th>
<th><strong>Dimensions</strong></th>
<th><strong>Records</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Columns First Name + Last Name vs. Age + ZIP Code</td>
<td>Symptoms Cough + Fever vs. Headache + Dangling Ankle</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Characteristics</strong></th>
<th><strong>Dimensions</strong></th>
<th><strong>Records</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Columns of Scale $[10^5...10^6]$ vs. Columns of Scale $[0..1] + [10^5...10^6]$</td>
<td>Undefined values vs. Defined values</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Statistics</strong></th>
<th><strong>Dimensions</strong></th>
<th><strong>Records</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlated Columns vs. Uncorrelated Columns</td>
<td>Records from the same cluster vs. Records from different clusters</td>
</tr>
</tbody>
</table>
II. Visual Analytics of Inhomogeneous Data

The Divide Step: 2-Step-Subdivision of Inhomogeneous Data

Note: Division does not need to be disjoint – a dimension can appear in multiple groups.
II. Visual Analytics of Inhomogeneous Data

Treat them differently:

VisBrick

VisBrick

VisBrick
II. Visual Analytics of Inhomogeneous Data

The Conquer Step: Layout

Multiform Visualization
II. Visual Analytics of Inhomogeneous Data

The Conquer Step: Linking
II. Visual Analytics of Inhomogeneous Data

The Result
II. Visual Analytics of Inhomogeneous Data

Current Research & Future Work: Extension to graph-structured data

Steffen Hadlak, Hans-Jörg Schulz, and Heidrun Schumann: In Situ Exploration of Large Dynamic Networks, appeared at InfoVis’11
II. Visual Analytics of Inhomogeneous Data

Current Research & Future Work: Extension to heterogeneous data?

Source: Richard Cyganiak, Anja Jentzsch
II. Visual Analytics of Inhomogeneous Data

Blatantly overreaching conjecture:
The Heterogeneity-Inhomogeneity-Duality (working title)

Data Federation/Warehousing/Fusion
(conquer)

Data Clustering/Partitioning
(divide)

HETEROGENEITY

INHOMOGENEITY
III. Orientation and Navigation in Heterogeneous Data

III. Orientation and Navigation in Heterogeneous Data

Data Heterogeneity
- multiple data sources
- which are linked via IDs, etc.
III. Orientation and Navigation in Heterogeneous Data

Data Heterogeneity
- multiple data sources
- which are linked via IDs, etc.

Visual Heterogeneity
III. Orientation and Navigation in Heterogeneous Data

Data Heterogeneity
- multiple data sources
- which are linked via IDs, etc.

Visual Heterogeneity
Analytical Heterogeneity
III. Orientation and Navigation in Heterogeneous Data

Orientation:

Where am I and where can I go from here?

Navigation:

Given a goal, which visual and/or analytical interface to use on which data set with which objective and in which order to reach this goal?

A typical goal is, for example:

treatment planning for cancer patients
III. Orientation and Navigation in Heterogeneous Data

Spell out the situation
Collect standard tasks

- X-Ray Images
- MR images
- CT images
- Tissue Samples
- Lab Results
- Anamnesis
- Patients' Info
- Diseases (ICD-10)
- Protein Expression
- mRNA Gene Expression
- Protein DB (NCBI)
- Gene DB (NCBI)
- Pathways (KEGG, BioCarta)
- Publications (PubMed)

View tissue
Inspect lab results
Record treatment decision
Determine patients
Browse patients
Select patients
Inspect images
Compare images
Segment tumor
Search disease
View disease description
Cluster expression data
Filter expression data
Find gene
View gene information
Explore related pathways
Find pathway
Investigate related articles
III. Orientation and Navigation in Heterogeneous Data

Spell out the situation

Collect standard tasks

Strip away their domain specificity…
III. Orientation and Navigation in Heterogeneous Data

Spell out the situation

Collect standard tasks

Strip away their domain specificity…

Link them to appropriate visual and analytical interfaces…

Diagram showing various data sources and analysis tools.
III. Orientation and Navigation in Heterogeneous Data

Spell out the situation
Collect standard tasks
Strip away their domain specificity...
Link them to appropriate visual and analytical interfaces...
Link them to the data sources they can be carried out on...
III. Orientation and Navigation in Heterogeneous Data

Model the Work Flow

- Use the collected standard tasks as building blocks for the work flow model
- Model alternative paths by branching out the work flow
- Use pre- and post-conditions to define the objective of a task
- ...

Note: The work flow is independent of the modeled setup!
Bring Work Flow and Setup together

- Determine exactly which visual or analytical interface to use for each operator of each step’s task
- Missing data sources and missing operators lead to a **pruning** of the workflow
- If the pruning doesn’t leave a path from start to finish, the goal cannot be achieved, but the smallest gap to close can be determined…
III. Orientation and Navigation in Heterogeneous Data

A few Words about the Cost of Modeling

- Mostly suitable for highly repetitive tasks in which each step is of utmost importance (airplane checks, biomedical procedures,...)

- Cost is not as high as imagined

1) hospitals often already have a data model for their IT infrastructure

2) workflows can be crowdsourced (unless they are proprietary)
III. Orientation and Navigation in Heterogeneous Data

Current Research & Future Work:

- Guidance across multiple users
- Guidance across applications

Source: Marc Streit, PhD Thesis (Graz, 2011)
III. Orientation and Navigation in Heterogeneous Data

Current Research & Future Work: Guidance across multiple displays

Source: Marc Streit, PhD Thesis (Graz, 2011)
IV. Conclusion, Food for Thought

My vision for heterogeneous data: Google Maps for Information Landscapes - to combine the meta view of the “data model” with multiform visualizations
IV. Conclusion, Food for Thought

My vision for heterogeneous data: Google Maps for Information Landscapes - to combine the meta view of the “data model” with multiform visualizations

Source: Richard Cyganiak, Anja Jentzsch
IV. Conclusion, Food for Thought

My vision for heterogeneous data: Google Maps for Information Landscapes - to combine the meta view of the “data model” with multiform visualizations

Source: Student Project by Grigor Tshagharyan (2011)
IV. Conclusion, Food for Thought

My vision for heterogeneous data: Google Maps for Information Landscapes - to combine the meta view of the “data model” with multiform visualizations
IV. Conclusion, Food for Thought

My vision for heterogeneous data: Google Maps for Information Landscapes - to combine the meta view of the “data model” with multiform visualizations
Acknowledgements

Further Information

Dr.-Ing. Hans-Jörg Schulz

Web: http://hjschulz.net
eMail: contact@hjschulz.net