Visual Analytics of Complex Data

Luke Zhang & John Yen
{lzhang, jyen}@ist.psu.edu

College of Information Sciences and Technology
The Pennsylvania State University
Problem We Try to Address

• Information overload in the analysis of complex data
• Limited tools to support in-depth analysis
  • Data processing algorithms to reveal hidden patterns
  • Visualization tools to analyze the patterns

• Combining user-centered interactive visualization and data-oriented computational algorithms
A Prototype on Engineering Design

Simulated Design Data

Projecting

User-centered Visualization

Scatter Plots

Model Building

Model Visualization

Data-oriented Analysis

Classification and Clustering

Parameter Refinement

Feedback Loop

User Interaction

Findings and Insights
Data Algorithms

• Decision Tree
  • Based on given data, what are those possible decisions?

• Data Clustering
  • Reduce data complexity and find common patterns
Design
Charts

- Data Presentation and Interaction

- Variable Matrix
Decision Tree
Clustering

- Hierarchical Clustering Visualization

\[ \sigma = 0.25 \]

\[ \sigma = 0.15 \]
Coordinated Views
Map-based Decision Making

Private Map

Public Map

Chatting Tool

Sorting Table

Aggregation Chart

Activity Timeline
Social Network Visualization
A Computational Social Science Approach

- Large scale online forum data (posts)
  - American Cancer Society’s Cancer Survivor Network (CSN)
  - The forum contains nearly half million posts.
- Automatic sentiment analysis
  - Train sentiment models using text mining and machine learning techniques
- Questions:
  - What are the impacts of online forum interactions to the sentiment of cancer survivors?
  - What factors are associated with the impacts?
CSN Forum Data

- Deidentified forum posts
  - From July 2000 to October 2010
  - 48,779 threads, 468,000 posts, 27,173 users

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of posts by a user</td>
<td>17.25</td>
<td>2</td>
<td>5,607</td>
</tr>
<tr>
<td>Number of replies per thread</td>
<td>8.7</td>
<td>6</td>
<td>442</td>
</tr>
<tr>
<td>Life span of a thread</td>
<td>1,725 hrs</td>
<td>58 hrs</td>
<td>87,846 hrs</td>
</tr>
</tbody>
</table>
Automatic Sentiment Analysis

- Model training
  - Manually labeled 298 posts
  - Feature extraction
  - Sentiment model learning and selection

- Model prediction
  - Apply the model to label the sentiment of un-labelled posts
  - Pr: the probability of a post to be positive
  - If Pr>0.5, Label=Positive, Otherwise, Label=Negative
Does the number of participants affect the sentiment change?

- The greater the number of replies by others to a thread, the more likely the thread originator’s subsequent sentiment is positive, regardless of the sentiment of his/her initial post.
92.1% of them have $\Delta_{pr} > 0$.

The average $\Delta_{pr}$ is 0.135, the standard deviation is 0.3

A t-test concludes that $\Delta_{pr}$ is significantly larger than 0
Potential Collaboration Opportunities

• Visual analytics of complex, multi-dimensional data
  – Discovering patterns hidden inside data
  – Helping to validating and verifying analytical models

• Visual analytics of public opinions
  – Using data-mining to identify opinion leaders (individuals and organizations)
  – Sensing the public sentiments towards business practices and new policies