Visually Analyzing A Billion Tweets: An Application for Collaborative Visual Analytics on Large High-Resolution Display

Simon Su, Michael An, Vincent Perry, Jianfeng Jia, Taewoo Kim, Te-Yu Chen, Chen Li
• 24 professional grade 42” LCD displays
• Slightly curved configuration of 6 columns by 4 rows
• Full HD resolution
• ~ 49.76 megapixel resolution across all 24 displays
• 4 machine cluster interconnected via 40 GbE Mellanox network
• Each runs Linux CentOS 7
• 1 head server node
  • hosts software packages / visualization server application
  • main file system
• 3 client display nodes
  • each contains 2 nVidia Quadro M6000s - 8 outputs
  • run motif window manager
  • use mosaic mode to run one full desktop across all 8 displays
- Enables use of high-resolution display system as multi-user workspace environment
- Web Server that runs on server machine
  - distributes rendering tasks to display clients
- Display clients run on display wall, one client per machine
  - connect to server via WebSockets
  - content data is managed on each node and appropriate portion of overall frame buffer is rendered
- UI clients can be connected through a normal browser
  - used to interact with wall
  - launch applications, drag/drop files
- Visual Analytic application
- Coordinated views over map and 2D bar chart
Large Scale Analytics System Components

Server
- AsterixDB
- Cloudberry Middleware

Client
- ParaViewWeb based client Application

Hardware Interface
- SAGE2 on Multi-tile Display
- Visualization Updates
- UI Interaction
- Query Input

Server-client interface
- Data Results
- Query

Human-Computer Interface
- Server-client interface
- Paraviewweb
- Provider / data modules enable coordinated data delivery across visualization components
  - “linking & dragging”
- View Components – given the information from the providers, renders the resulting visualizations
- Requests processed in CloudberryQueryModel, and sent out to the Cloudberry middleware
- Data received from Cloudberry, flows modularly through the providers, and to the view components
- Cache previously run queries for quick access
- Deliver data in a progressive manner
  - Using WebSocket connection
INFORMATION SCALING - AMOUNT OF INFORMATION DISPLAYED AT ONE TIME

VISUAL SCALING - AMOUNT OF INFORMATION DISPLAYED AT HIGHER RESOLUTION

LHRD

Normal

LHRD

Normal
SERIES OF IDENTIFICATION AND COUNTING TASKS COMPLETED BY PARTICIPANT
ACCURACY AND SPEED OF TASKS RECORDED
SURVEY QUESTIONS ASKED AFTER TASKS COMPLETED
<table>
<thead>
<tr>
<th>ID</th>
<th>Environment</th>
<th>Task Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Map</td>
<td>Select the 5 counties with highest tweet counts</td>
</tr>
<tr>
<td>2</td>
<td>Map</td>
<td>Select the 3 countries with lowest tweet counts greater than 0</td>
</tr>
<tr>
<td>3</td>
<td>Histogram</td>
<td>Select the 3 histograms with the most area</td>
</tr>
<tr>
<td>4</td>
<td>Combined</td>
<td>Select the 5 counties with the highest tweet counts</td>
</tr>
<tr>
<td>5</td>
<td>Combined</td>
<td>Count the number of regions that were colored by any color</td>
</tr>
</tbody>
</table>
Time to completion of tasks performed on normal and large displays
1. **IN WHICH DISPLAY SIZE DID YOU MORE QUICKLY FIND WHAT YOU WERE LOOKING FOR?**

2. **IN WHICH DISPLAY SIZE DO YOU FEEL YOU FOUND MORE OF THE PATTERNS OF INTEREST AND LESS OFTEN MISSED IMPORTANT PIECES OF THE VISUALIZATION?**

3. **IN WHICH MODE WAS UI INTERACTION BETWEEN THE HISTOGRAMS AND MAP MORE INTUITIVE?**

4. **IN WHICH MODE WAS IT EASIER TO UNDERSTAND THE BIG PICTURE?**

5. **WHICH MODE ALLOWED FOR GREATER UNDERSTANDING OF FINE DETAIL?**

6. **WHICH MODE DID YOU PREFER?**
FREE RESPONSE FOR USERS TO PROVIDE COMMENTS WHY THEY PREFERRED LHRD OVER NORMAL DISPLAY:
Ability to see everything at once without scrolling
Easier to compare amongst map regions
More granular detail making it easier to discern between counties
Easier to understand/see all the data
TASKS ON LHRD ACTUALLY TOOK LONGER, BUT USERS ANSWERED THEY FELT THAT NORMAL TOOK LONGER.

PROMISING THAT USERS ANSWERED THIS WAY, BUT WHY IS IT TRUE?

SPECULATION:

On LHRD, all data can be considered at once and in comparison to all other data – takes longer to analyze.

On normal, only subset of data is considered at once, causing a quicker response from users.

LHRD responses may be more accurate due to observability of all data.

FUTURE: QUANTIFY "ACCURATE" RESPONSES FOR TESTING.
Questions

simon.m.su.civ@mail.mil