Overview of the Visual Analytics Challenge

Visual Analytics Evaluation Workshop
May 29th, 2009 – University of Maryland

Catherine Plaisant

with Jean Scholtz, Georges Grinstein and Mark Whiting (VAST Challenge co-chairs)
Swetha Reddy and Loura Costello (Grad students)
Heather Byrne, Adem Albayrak (NSF REU undergrad)
Visual Analytics Tools

Analytics algorithms
Interactive interfaces
Visualizations
Reasoning tools
Reporting tools
Collaboration
Many application areas

- **Financial analysis**
  - Is this drug money laundering?
- **Public health**
  - Should we fear this swine flu epidemics?
- **Intelligence analysis**
  - Is this dogster behavior suspicious?
- **Business analysis**
  - What are my competitors doing?

Analyze data, support or refute hypotheses
Report findings, recommend actions
Who needs to evaluate the utility of the tools?

• Developers
  – Does it work?
  – How to improve?

• End-users = Analysts
  – How to understand and select tools?
  – How to encourage better designs?

• Sponsors
  – Is this a good use of my money?
  – What areas needs more research?

• ALL: What’s missing?
Levels of Evaluation

Example Metrics
- Adoption
- Productivity
- Satisfaction
- Utility
- Learnability

Complex Processes
Simpler Tasks

User-Centered Metrics

Work Environment

System
- Human-Information
- Discourse Methods
- Collaboration

Components
- Interface Designs
- Interaction Techniques
- Visual Representations
- Analytics Algorithms

From Plaisant & Laskowski, in Evaluation section of Illuminating the Path
Evaluation methods

• **Usability studies**
• **Controlled experiments**

  - Too short and simple tasks!

• **Insight studies**
  – Open ended exploration for a few hours
  – Measure insights gained with tools being compared

  ---- still too short, hard to conduct

• **Longitudinal studies**
  – One tool, many uses
  – Better, but hard to compare tools!
Evaluation methods

• Usability studies
• Controlled experiments
  
  --- Too short and simple tasks!

• Insight studies
  – Open ended exploration for a few hours
  – Measure insights gained with tools being compared

  ---- still too short, hard to conduct

• Longitudinal studies
  – One tool, many uses
  – Better, but hard to compare tools!
Evaluation methods

• Usability studies
• Controlled experiments

  - - - Too short and simple tasks!

• Insight studies
  – Open ended exploration for a few hours
  – Measure insights gained with tools being compared

  ---- still too short, hard to conduct

• Longitudinal studies
  – One tool, many uses
  – Better, but hard to compare tools!
Evaluation methods

• Usability studies
• Controlled experiments
  
   - - - Too short and simple tasks!

• Insight studies
  – Open ended exploration for a few hours
  – Measure insights gained with tools being compared

    ---- still too short, hard to conduct

• Longitudinal studies
  – One tool, many uses
  – Better, but hard to compare tools!
Longitudinal studies

- Ethnographic Studies
- Milcs
  - Multidimensional In-depth Longitudinal Case Studies (Beliv'06)
  - Goal is to document discoveries
  - Working with users and refining tools

Professional weather forecasters

Counter terrorism researcher - Understanding the Global Jihad terrorist Network

Political analyst - U.S. Senators voting patterns

e.g. Trafton et al, 2000
Professional weather forecasters
Evaluation methods

• Usability studies
• Controlled experiments
  - - - Too short and simple tasks!

• Insight studies
  – Open ended exploration for a few hours
  – Measure insights gained with tools being compared

  ---- still too short, hard to conduct

• Longitudinal studies
  – One tool, many uses
  – Better, but hard to compare tools!

And many more methods...
Software testing
Requirement traceability
Surveys
User workshops
Log analysis
Glassbox
etc.
Evaluation challenge

• Long complex tasks
• Collaborative
• Integrate different type of data
• Use multiple tools over a long time
• Context and tacit knowledge important
• Scalability is an big issue

• But researchers often have
  Limited access to data and users
Contests

- Realistic scenarios and datasets
- Months to conduct the tasks

- Rewards

- Venue to discuss evaluation

- Improvement of evaluation methodology
  - Metrics
  - Evaluation process

Case based evaluation
InfoVis 2004 Contest
10 years of Infovis papers
www.cs.umd.edu/hcil/iv04contest

No ground truth
No accuracy measure
Still no way to assess utility
The VAST Challenges (2006-2009)

• Invented scenario and synthetic datasets (with ground truth that only us know) developed at PNNL by Mark Whiting team

• Combine accuracy ratings and subjective assessment
2006/2007 Challenge data and tasks

- Mostly text
- Whodunnit
- 7-10 entries
- Many submissions not clearly explained
- Lots of interest at VAST symposium
- Participants reported clear benefit

- ??? How to increase participation ????
VAST 2008 Challenge data and tasks

– Background info
  about Paraiso movement - history and beliefs of the group

– 4 Mini-Challenge datasets
  – 10 days of Cell phone calls (characterize change in network)
  – 3 years of Migrant Boats landings (characterize change)
  – Evacuation traces (identify suspects after explosion)
  – Wiki page edits (describe the factions)
    (Teams may enter one or more)

– Grand Challenge integrates all 4
  – Assess beliefs of the movement and their activities
  – Determine if the movement advocates violence
VAST 2008 Challenge data and tasks

– Background info
  about Paraiso movement - history and beliefs of the group

– 4 Mini-Challenge datasets
  – 10 days of Cell phone calls (characterize change in network)
  – 3 years of Migrant Boats landings (characterize change)
  – Evacuation traces (identify suspects after explosion)
  – Wiki page edits (describe the factions)
  (Teams may enter one or more)

– Grand Challenge integrates all 4
  – Assess beliefs of the movement and their activities
  – Determine if the movement advocates violence
Teams go to work....

Months to participate and submit:
- Answers
- Video demonstration
- Process description

Strong interest
- 73 submissions
- 6 Grand Challenge entries
- 28 organizations
- 12 student teams
- 13 countries
Judging

- Judges
  - Interface and visualization experts
  - Professional analysts

- Criteria
  - Accuracy of the answers
  - Subjective assessment of utility of tools in arriving at the answers

(Can talk a lot more about that…)
Examples of Awards

• Analysis
  ■ Accuracy
  ■ High quality report

• Visualization
  ■ Intuitive visualizations
  ■ Innovative visualizations

• System
  ■ Outstanding functionality
  ■ Level of integration
Examples from Migrant Boat Mini Challenge

- **Dataset**
  - 3 years log of landings and interdictions by Coast Guard (= arrests)
Examples from Migrant Boat Mini Challenge

• Dataset
  – 3 years log of landings and interdictions by Coast Guard (= arrests)

• Question
  “Characterize - choice of landing sites
  - patterns of interdiction and evolution over 3 years”
Solution (Analytic Situation)

(partial)

• Landing strategy moved *westward* until successful landing began in 2007 in Yucatan Peninsula, Mexico.
Diversity of approaches: e.g. Symbol used on the map

- Dots - CORE
- Lines - Tacc
- Centroid - Tacc
- SPADAC - Arrows
Aggregation – or not...
Aggregation – or not…

Entries without aggregation missed small island

<< || > >>

replay - animation only
Aggregation – or not…

Parvac – U of Washington
L5 = Mexico

Oculus
Irregular regions
Aggregation shown on map
Automatic Grouping and Clustering

How labeled?
Connection to map?
Not explained!

Can MY TEAM do better?
Representing Time

- Animation
- Timelines
Representing Time

- 3D

- Hybrid Projection
Quality of analysis

• Many merely report yearly counts
• Some made hypotheses
• Few provided:
  - analysis support
  - reporting support

Oculus

<table>
<thead>
<tr>
<th>Year</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>276</td>
</tr>
<tr>
<td>2006</td>
<td>301</td>
</tr>
<tr>
<td>2007</td>
<td>465</td>
</tr>
</tbody>
</table>
Quality of analysis

• Many merely report yearly counts
• Some made hypotheses
• Few provided:
  - analysis support
  - reporting support

Teams learn from each others

Oculus

2005 = 276
2006 = 301
2007 = 465
Analysis thoroughness

• Minimum answers
• Or looked at data very thoroughly
• Or made suggestions for action

Ground truth grows with # of submissions
→ Rapid verification by analysts
→ Dynamic aspect of ratings
Cell Phone Logs Mini Challenge

Ground truth included:

“On day 7,
Paraiso group dropped their phones and picked up another set of phones”.

Many teams did not see it!
Some did, either:
- entirely visually
- using numerical metrics
  (high Standard Deviation of eigenvector centrality)
Entire network

Paraiso group

From Palantir submission
• Collaboration: No award!

• Scalability: No Award!

No award = more research needed
Award for move in the right direction
e.g. Data Integration

Pennsylvania State University-NEVAC
Award for Innovative Visualization

Southern Illinois University Edwardsville
Award for Interactive Visual Analytic Environment

Palantir Technologies
Award for Analysis Summary (debrief)

“Mexico landings could be seen as a land bridge to Florida” ../..

“these were successful because US Coast Guard would have no jurisdiction in Mexico.”

SPADAC Inc.
Motivation to participate
Motivation to participate

• Publication, award for resume
• Visibility

• Participate in Challenge Workshop
• Benefits
  – Increased awareness of need for evaluation
  – Repository of data + tasks + examples of use
  – Improved methodology
    • Accuracy metrics
    • Subjective assessment from Experts AND Professional Analysts

• Limitations
  – Mostly “whodonnit” scenarios
  – (Still) Too small and simple
  – Custom automatic evaluation
  – Doesn’t address wide diversity of domains
  – No follow-up
  – Etc. etc.
Lesson learned?

• Mini-challenges → increased participation
• Automatic evaluation a desirable goal but subjective assessment from experts and analysts will always remain crucial
• Dataset generation is time consuming
• Student involvement was significant
• A learning experience for everyone
• No ranking but rewarding of best designs
Possible ingredient for success

- Continuity of Challenge committee team (we learn too!)
- Publishing of materials
  - Submissions improve over the years from past examples
- Variety of reward mechanisms
- Support
  - for developing datasets
    - topics accessible to everyone
    - clever scenarios to motivate teams
  - for managing event
  - for analyst time
Still: sustainability issue…
Ransom of success…

• Reviewing logistics issues
  – In 2009: need triage of mini challenge entries by external reviewers
  – Development of submission and review website
    Multimedia review system + computing of accuracy measures + teams with multiple submissions

• Resource strain
  – Demand for diverse topics (instead of refinement of one topic)
  – Judging in other venues (e.g. KDD)

• Tension between service and research
  – Development of websites and services needed
Conclusions

• Participate and learn from others!
• Make scenarios and data available
• Support evaluation activities

plaisant@cs.umd.edu
www.cs.umd.edu/hcil/VASTchallenge09

Thanks to Jean Scholtz, Georges Grinstein and Mark Whiting (co-chairs) Swetha Reddy and Loura Costello (Grad students) Heather Byrne, Adem Albayrak (NSF REU undergrad) NSF, NVAC, NIST-IARPA