Hands-On: Visualization

Tamara Munzner
Department of Computer Science
University of British Columbia
IRS Cross-Border Reporting Workshop
May 2017, Vancouver BC

http://www.cs.ubc.ca/~tmm/talks.html#cb17
@tamaramunzner
Visualization (vis) defined & motivated

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

• human in the loop needs the details
  – doesn't know exactly what questions to ask in advance
  – exploration: longterm exploratory data analysis (EDA)
  – presentation: known results
  – stepping stone towards automation: refining, trustbuilding
• external representation: perception vs cognition
• intended task, measurable definitions of effectiveness

more at:
Visualization Analysis and Design, Chapter 1.
Two vignettes

• a tale of two tools created for journalistic use
  – shared frameworks of interdisciplinary methods from my research group
    • thinking about collaboration
      – roles & rewards, for computer scientists & journalists
    • reasoning about visualization design
      – beyond pretty pictures
  – divergent goals & audiences
    • TimeLineCurator: presentation / explanatory
    • Overview: investigation / exploratory
Collaboration incentives

• why do CS/vis people need to understand journalism’s problems?
  – we work with you to understand your driving problems
  – we build tools intended to help
    • only works out if we understood the problems deeply enough
  – we observe how you use them
    • if they’re good enough
      – CS win: research success stories
      – journalist win: access to better tools
  – we develop guidelines on how to build better tools in general
    • CS win: research progress in visualization
Vignette 1:
Vis Tool for Journalistic Presentation
Interactive Authoring of Visual Timelines from Unstructured Text

http://about.timelinecurator.org

http://timelinecurator.org

TimelineCurator: Interactive Authoring of Visual Timelines from Unstructured Text.
Origin story: Tedium in the newsroom

• Johanna Fulda: interactive infographics developer, Sueddeutsche Zeitung
  – then Munich CS master’s student, visiting UBC

• what pain point could we address with interactive visualization?
  – plus some natural language processing (NLP)
Manual creation process
Structured creation process

TimelineJS
timeline.knightlab.com/
Timeline authoring model

- time required for each task

<table>
<thead>
<tr>
<th></th>
<th>Browse</th>
<th>Extract</th>
<th>Format</th>
<th>Show</th>
<th>Update</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manual Drawing</strong></td>
<td>slow</td>
<td>slow</td>
<td>slow</td>
<td>slow</td>
<td>slow</td>
</tr>
<tr>
<td><strong>Structured Creation</strong></td>
<td>slow</td>
<td>slow</td>
<td>slow</td>
<td>automated</td>
<td>fast</td>
</tr>
<tr>
<td><strong>TimeLine Curator</strong></td>
<td>fast</td>
<td>automated</td>
<td>automated</td>
<td>fast</td>
<td>fast</td>
</tr>
</tbody>
</table>
The general case for curation

- build for human in the loop as continuing need
  - automatic processing to accelerate not replace
  - assume computational results good but not perfect
    - for the indefinite future!
  - visual feedback to accelerate
The importance of being brisk

• sexy use case: eureka moment
  – enable what was impossible before
  – vis tools for new insights & discoveries

• workhorse use case: workflow speedup
  – vis tools to accelerate what you’re already doing
    • sometimes enables the previously infeasible

• TLC use cases
  – started with speedup use case, for presentation
    • make this doc into a timeline now!
  – two other use cases nudge towards exploration
    • comparison between multiple timelines
    • speculative browsing
TimeLineCurator: Speculative Browsing

https://vimeo.com/jofu/tlc
Try it out!

http://timelinecurator.org
Vignette 2: Vis Tool for Investigative Reporting
The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists

http://www.cs.ubc.ca/labs/imager/tr/2014/Overview/

https://www.overviewdocs.com

Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool For Investigative Journalists.
Starting point: Dimensionality reduction for document datasets

Task 1
In HD data → Out 2D data

What?
⇒ In High dimensional data
⇒ Out 2D Data

Why?
⇒ Produce
⇒ Derive

Task 2
In 2D data → Out Scatterplot Clusters & points

What?
⇒ In 2D data
⇒ Out Scatterplot
⇒ Out Clusters & points

Why?
⇒ Discover
⇒ Explore
⇒ Identify

How?
⇒ Encode
⇒ Navigate
⇒ Select

Task 3
In Scatterplot Clusters & points → Out Labels for clusters

What?
⇒ In Scatterplot
⇒ In Clusters & points
⇒ Out Labels for clusters

Why?
⇒ Produce
⇒ Annotate

• more on DR: hour-long talk Dimensionality Reduction from Several Angles
  http://www.cs.ubc.ca/~tmm/talks.html#kelowna16
Origin story: WikiLeaks meets Glimmer

• WikiLeaks: hacker-journalist Jonathan Stray analyzing Iraq warlogs
  – one instance of general problem: Too Many Documents
  – conjectured that existing label classification falls short of showing all meaningful structure in data
    • friendly action, criminal incident, ...
  – he had some NLP, needed better vis tools

• Glimmer: multilevel dimensionality reduction algorithm
  – scalability to 30K documents and terms

[Glimmer: Multilevel MDS on the GPU.
Overview: Early version

http://www.cs.ubc.ca/labs/imager/tr/2012/modiscotag
## Deploy in the real world

<table>
<thead>
<tr>
<th>Case Study</th>
<th>Document Collection</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>4,500 pages from FOIA</td>
<td>What did security contractors do during Iraq war?</td>
</tr>
<tr>
<td>#2</td>
<td>5,996 emails from FOIA</td>
<td>Were municipal police funds mismanaged?</td>
</tr>
<tr>
<td>#3</td>
<td>8,680 pages from FOIA</td>
<td>Were Paul Ryan’s campaign statements hypocritical?</td>
</tr>
<tr>
<td>#4</td>
<td>1,278 survey comments</td>
<td>What is the gun ownership debate about?</td>
</tr>
<tr>
<td>#5</td>
<td>4,653 emails from FOIA</td>
<td>Was gov’t response to emergency incident effective?</td>
</tr>
<tr>
<td>#6</td>
<td>1,680 bills</td>
<td>Did gov’t fail to pass bills addressing police misconduct?</td>
</tr>
</tbody>
</table>
## Deploy in the real world, understand user goals

<table>
<thead>
<tr>
<th>Case Study</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Document Collection</strong></td>
<td>4,500 pages from FOIA</td>
<td>5,996 emails from FOIA</td>
<td>8,680 pages from FOIA</td>
<td>1,278 survey comments</td>
<td>4,653 emails from FOIA</td>
<td>1,680 bills</td>
</tr>
<tr>
<td><strong>Question</strong></td>
<td><strong>What did security contractors do during Iraq war?</strong></td>
<td><strong>Were municipal police funds mismanaged?</strong></td>
<td><strong>Were Paul Ryan’s campaign statements hypocritical?</strong></td>
<td><strong>What is the gun ownership debate about?</strong></td>
<td><strong>Was gov’t response to emergency incident effective?</strong></td>
<td><strong>Did gov’t fail to pass bills addressing police misconduct?</strong></td>
</tr>
<tr>
<td><strong>the trend story:</strong></td>
<td>get the big picture</td>
<td>find the needle in the haystack / smoking gun</td>
<td>wheat vs chaff: filtering out irrelevant material</td>
<td>categorize and count: turning docs into data</td>
<td>wheat vs chaff: filtering out irrelevant material</td>
<td>prove haystack contains no needles!</td>
</tr>
</tbody>
</table>

Evolution across levels

• evolution of task abstraction
  – task 1: generate hypotheses → explore → summarize
    • obviously you can’t read everything; speed up with tool for categorizing and counting
  – task 2: verify hypotheses → locate → identify
    • you really do read each doc; speed up with tool to keep track of findings

• evolution of data abstraction & idioms
  – arrange cluster tree to emphasize nodes vs links
  – new vis insight: DR scatterplot less effective than cluster tree vis + tagging
    • better affordance for systematic traversal of document cluster hierarchy

early

current
Try it out!

https://www.overviewdocs.com/

https://blog.overviewdocs.com/completed-stories/
Discussion

• how many of you have jumped into data journalism?
  – what kind of tools have you used?

• have any of you grappled with timelines?
  – what kind of tools did you use?

• have any of you grappled with large document collections?
  – what kind of story did you have in mind?
More Information

• this talk
  www.cs.ubc.ca/~tmm/talks.html#cb17

• book
  http://www.cs.ubc.ca/~tmm/vadbook
  – 20% off promo code, book+ebook combo: HVN17

• papers, videos, software, talks, courses
  http://www.cs.ubc.ca/group/infovis
  http://www.cs.ubc.ca/~tmm