Data Visualization (CIS/DSC 468)

Text Visualization

Dr. David Koop
Fields in Visualization

Scalar Fields
(Order-0 Tensor Fields)

Vector Fields
(Order-1 Tensor Fields)

Tensor Fields
(Order-2+)

Each point in space has an associated...

Scalar

\[ s_0 \]

Vector

\[
\begin{bmatrix}
  v_0 \\
  v_1 \\
  v_2 
\end{bmatrix}
\]

Tensor

\[
\begin{bmatrix}
  \sigma_{00} & \sigma_{01} & \sigma_{02} \\
  \sigma_{10} & \sigma_{11} & \sigma_{12} \\
  \sigma_{20} & \sigma_{21} & \sigma_{22} 
\end{bmatrix}
\]
Examples of Vector Fields

Wind [earth.nullschool.net, 2014]
Examples of Vector Fields

Wind [earth.nullschool.net, 2014]
Rendering Vector Field Statistics as Scalars

• Many statistics we can compute for vector fields:
  - Magnitude
  - Vorticity
  - Curvature

• These are scalars, can color with our scalar field visualization techniques (e.g. volume rendering)

[Color indicates vector magnitude]
Glyphs

- Represent each vector with a symbol
- Hedgehogs are primitive glyphs (glyph is a line)
- Glyphs that show direction and/or magnitude can convey more information
- If we have a separate scalar value, how might we encode that?
- Clutter issues
Streamlines (Step 1)

\[ [x, y] \rightarrow [-y, (1/2)x], \text{ Step: 0.5} \]
Streamlines (Step 19)

\[ [x, y] \rightarrow [-y, (1/2)x], \text{ Step: 0.5} \]
Higher-Order Comparison

Euler vs. Runge-Kutta

- RK-4: pays off only with complex flows

Approx. like RK-2

Higher-Order Comparison

[via Levine]
Line Integral Convolution

- Goal: provide a global view of a steady vector field while avoiding issues with clutter, seeds, etc.
- Remember convolution?
- Start with random noise texture
- Smear according to the vector field
- Need structured data

[Weiskopf/Machiraju/Möller]
2D Scalar Field Topology
2D Scalar Field Topology

[Wikipedia]
2D Scalar Field Topology

Key development in topological data analysis (TDA)
1. Abstraction of the data: topological structures and their combinatorial representations
2. Separate features from noise: persistent homology

Reeb Graph/Contour Tree/Merge Tree
Vector Field Topology

• Instead of “guessing” correct seed points for streamlines to understand the field, try to identify structure (topology) of the field.

Figure 7.1 A phase portrait.

[M.Henle]
Topological Skeleton
Assignment 4

• Due Thursday Monday
• Changing value + reordering interaction
• Brushing (linked highlighting)
Assignment 5

- Due at the end of the semester
- Use ParaView for sci vis
  - Isosurfaces
  - Volume Rendering
  - Streamlines
  - Glyphs
- Mac Users: Download 5.2.0
- Turn in screenshot and state file
Schedule

• Monday, April 24: Guest Lecture: Prof. Sun (Sets)
• Wednesday, April 26: Sets & Clusters
• Friday, April 28: CIS Day, No Class
• Monday, May 1: Review
• Tuesday, May 9: Final Exam: 11:30am-2:30pm (Dion 101)
Text Visualization

• Why visualize text? Text is already visual, right?

• How much text? What granularity? (What is an item?)
  - Single string
  - Words/lines
  - One document
  - Multiple documents (corpus)

• Considerations:
  - Legibility
  - Variable length
  - Locality
  - Occurrence
  - Semantics
Data Sources

• Literature: books, poetry
• Social Media: tweets, posts
• Web: Pages, posts, emails
• Code
Tag Cloud

• Derived data: number of occurrences of words
• Channel: Font size
• Potential problem: Think about ink…
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had had no pictures or conversations in it, and what was the use of a book, thought Alice without pictures or conversation?

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

There was nothing so VERY remarkable in that; nor did Alice think it so VERY much out of the way to hear the Rabbit say to itself, ‘Oh dear! Oh dear! I shall be late!’ (when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural; but when the Rabbit actually TOOK A WATCH OUT OF ITS WAISTCOAT-POCKET, AND SET IT TO AT once, she felt it was time to say something.)
TextArc

• Three rules:
  - Show the entire text in an ellipse around the page: line-by-line and word-by-word
  - Like tag clouds, use larger font-size and brighter text for frequent words
  - Central words move to the middle (links to its mentions)
on Many Eyes, for instance, we would not have guessed at the popularity of religious analyses. Given the broad demand for text visualizations, however, it seems like a fruitful area of study.

ACKNOWLEDGEMENTS

The authors thank Frank van Ham, Jesse Kriss, Matt McKeon, Lee Byron, and Eric Gilbert for helpful suggestions. In addition, we are grateful to the users of Many Eyes for their creativity and willingness to provide feedback on an experimental visualization technique.

Fig 9. Word tree of the King James Bible showing all occurrences of “love the.”

Fig 10: Word Tree showing all occurrences of “I have a dream” in Martin Luther King’s historical speech.

[Wattenberg & Viegas, 2007]
Word Tree

• A "Visual Concordance"
• Shows phrasing, relationships between words
• Starting point is a single word or snippet
• Branches to show common words/phrases that follow
• Goal is to show context: "keyword-in-context"
Interaction in Word Tree

As with all visualizations on Many Eyes, users can set parameters for the display of the word tree. Many other interactive features are available, such as the ability to select particular words, by order of first occurrence in the text (the default option, although this may not always be the best choice, as it can cause the visualization to recenter to the longest phrase), by frequency (so the largest branches are those words that appear in the text most often), or by suffix-tree-like beginnings were attempted, but seemed too uninformative. A future version might try to automatically find a good starting point: perhaps a tree centered on the most frequent terms, with the deepest branches. Having a default start point saves a significant amount of memory; for instance, in the King James Bible, probably one of the most-visualized text on Many Eyes. (In figure A, this is the approximate number of tokens in the King James Bible, probably one of the most-visualized text on Many Eyes.) As the user interacts with the tree—she may click on a branch to “highlight” words to start a scan for additional phrases after this term. In C, the user Control-clicks on “blind,” which appears in one of the branches under “if.” This causes the visualization to recenter to the longer phrase “if love be blind,” which causes the visualization to recenter to the longer phrase “if love be blind.” In B, the user is given an overview. The reason for this design choice is that without a search term, there is no obvious entry point—several alternatives are possible. We discuss these in the sections on user feedback and future work.
On the other hand, applying different expressions to the same text can reveal a series of interrelated conceptual networks. The phrase nets of Jane Austen's novel *Pride and Prejudice* in Fig. 5 illustrate this. Matching "X and Y" shows a network of concepts and people. The main characters appear neatly organized in two clusters: Jane, Elizabeth, Lydia, Kitty, Catherine and Mr. Bingley form a central cluster, whereas "mother," "aunt," and "uncle" keep some distance. Positive attributes such as "sense," "disposition," "humour," "kindness" cluster together while less flattering qualities such as "pride," "conceit," "vanity," "folly," and "ignorance" form a group of their own. Perhaps most interesting, to those familiar with the novel, is that "Darcy" does not appear in the network—in a certain sense he is the most solitary major character.

If we analyze the same text with the pattern "X at Y" we obtain an entirely different network that reveals the set of locations inhabited by the characters in the novel and the events that take place at those locations. In a sense, the user can direct exploration towards a particular dimension of the text by intelligently choosing the pattern to match for.

Figure 1 shows the result of another targeted pattern. Here we have analyzed the whole bible using the pattern "X begat Y," a specific formulation from the King James Bible indicating a parent-child relationship. The resulting graph illustrates the lengthy genealogies that are recorded by many different books in the bible. The network also uncovers a number of defining aspects of these lineages, such as the importance of Abraham.

### 4.2 Regular expressions and matching
The patterns we have shown so far are of the form "X <connector> Y", where the connector is either a separate word or a phrase. However, regular expressions also allow us to specify patterns that match for specific pre- and postfixes to X and Y. Previously, the authors worked with a humanities scholar to analyze a set of 7,000 British novel titles between 1740 and 1850—In fact, much of the motivation behind building phrase net comes from this collaboration. This scholar was interested in how the use of simple syntactic constructions such as "X of the Y" reflected changes in literary style over the centuries.

---

**Fig 5.** Matching different patterns on the same text. Here we analyzed Jane Austen's *Pride and Prejudice* with "X and Y" and "X at Y" respectively. The left image shows relationships between the main characters amongst others, while the right image shows relationships between locations.

**Fig 4.** Matching the same pattern on different texts. Here we used the pattern "X of Y" to compare the old and new testaments. Israel takes a central place in the Old Testament, while God acts as the main pattern receiver in the New Testament.

---

**Phrase Nets**

*van Ham, 2009*
Words are more than just character sequences

Fed Drapes
Clark Coolidge

FELL FAR BUT THE BARN (came) up & smacked me
Who’re you, bleeding? Fed.
Blat in back of a Vistrola car
is so red is such that sun
fell in the rushes & pen bear appear

the white wrong numeral on the wall
can’t take if off with the clock
down with the clock it...
way
on the board couch with brass, kindergarten clench joints
backed violet rip into the gas valve
it hemmed & snowed

the wrong way
remnant face
rubber
the pucker

Rhymes | Phonetic Rhymes | Character Clusters | Levenshtein Distance
---|---|---|---
Identical Rhyme/Rhyme Riche
Perfect Rhyme
Semirhyme
Syllabic Rhyme
Consonant Slant Rhyme
Vowel Slant Rhyme
Pararhyme
Eye Rhyme
Alliteration
Assonance

[N. McCurdy et al., 2015]
Poemage

• Support close reading—in-depth reading to generate as much productive meaning as possible
• Search for poetic devices: affect, imagery, pun, metaphor
• Sound and linguistic devices → Rhyming
  - Identical: pare/pair
  - Perfect: picky/tricky
  - Assonance & consonance: blue/estuaries, shell/chiffon
  - Eye rhyme: cough/bough
• Support exploration: scholars do not want computers to "solve" poems

[N. McCurdy et al., 2015]
Poemage: Visualizing the Sonic Topology of a Poem

Nina McCurdy, Julie Lein, Katharine Coles, Miriah Meyer

Fig. 1. The Poemage interface comprises three linked views: (left) the set view allows users to browse sets of words linked through sonic and linguistic resemblances; (middle) the poem view allows users to explore sonically linked words directly via the text; (right) the path view shows the sonic topology of a poem.

Abstract

The digital humanities have experienced tremendous growth within the last decade, mostly in the context of developing computational tools that support what is called distant reading—collecting and analyzing huge amounts of textual data for synoptic evaluation. On the other end of the spectrum is a practice at the heart of the traditional humanities, close reading—the careful, in-depth analysis of a single text in order to extract, engage, and even generate as much productive meaning as possible. The true value of computation to close reading is still very much an open question. During a two-year design study, we explored this question with several poetry scholars, focusing on an investigation of sound and linguistic devices in poetry. The contributions of our design study include a problem characterization and data abstraction of the use of sound in poetry as well as Poemage, a visualization tool for interactively exploring the sonic topology of a poem. The design of Poemage is grounded in the evaluation of a series of technology probes we deployed to our poetry collaborators, and we validate the final design with several case studies that illustrate the disruptive impact technology can have on poetry scholarship. Finally, we also contribute a reflection on the challenges we faced conducting visualization research in literary studies.

Index Terms

Visualization in the humanities, design studies, text and document data, graph/network data

1 INTRODUCTION

The use of digital tools across disciplines in the humanities has exploded during the last decade. Popular projects such as the Google Ngram Viewer [37] and Wordle [55] have harnessed the power of computation to look across huge corpora of texts, leading to insights that had never been available before. Tools such as these are highly effective in supporting what is called distant reading—a term coined by literary scholar Franco Moretti to describe critical approaches that seek to understand literature and literary history by aggregating and quantitatively analyzing large text corpora.

Despite this new mode of scholarship, traditional humanities scholars continue to engage primarily in a very different type of analysis called close reading. As its name implies, close reading involves a detailed analysis of a text in all its complexity, encompassing an analysis not only of specific operations such as syntax, rhyme, and meter; such figures as metaphor and allusion; and such linguistic effects as affect, but also of how these operations interact across the temporal and spatial field of the text, with each other and with the reader, to create meanings greater than the sum of the parts. As this description suggests, much of the work done in close reading is well beyond the current capabilities of computation. Thus, the true value of computation to close reading is still very much in question and is the topic of an ongoing dialogue in the digital humanities. While a handful of computational tools have been designed to support close reading, much of the problem space remains unexplored.

We conducted a two-year design study with poetry scholars and practitioners to explore this gap. Our two primary collaborators, both of whom are co-authors on this paper, identify both as poets and as academics. We also engaged a network of practitioners, including two professors and two students of poetry. Together, these collaborators have literary expertise in medieval, early modern, modernist, and contemporary poetry, and they analyze poetry from a range of traditions and periods. Furthermore, they write formal verse, free verse, and experimental poems, and thus bring a diversity of theoretical viewpoints to their critical and creative work.

During this design study, we encountered several specific challenges that affected our design process. First, supporting close reading...
Comparing Documents

- Word choice/usage
- Relationships
- Phrasing
Tag Cloud (Two Documents)

State of the Union Address, 2002 vs. 2011

President Bush, January 29, 2002

President Obama, January 25, 2011

[Pyrsmis, CC-BY-SA-3.0]
Figure 1: A PTC revealing the differences in drug prevalence amongst the circuits. ally similar to the connected lists view of Jigsaw [28], however PTCs use size-weighting of words in the display. Shneiderman and Aris [26] have previously explored the contents of a faceted legal document databases using matrix-based visualizations to reveal the number and type of data items matching each facet value. Our work differs in that we seek to aggregate and visualize the contents of the data items, not only their presence or absence. A matrix visualization approach would not be appropriate as our word-selection method, described later, seeks to maximize the differences between corpus subsets. Rather than the single vertical column of words that a words \( \times \) facets matrix would contain, our approach allows the entire space to be filled with a wide variety of words. VisGets, or visualization widgets, have been used to explore faceted collections of web-based streaming data [5]. Facets are filtered using scented visual widgets [34] appropriate for the data type, providing both an overview of the available data items and a method to drill down along several facets simultaneously. A tag cloud VisGet consists of a traditional tag cloud summarizing all available documents — text differentiation along a facet is only achieved through interactive brushing. The goal of VisGets is to provide coordinated overview and navigation tools in a faceted information space, where our work is customized to providing meaningful differentiating overviews across facets within large amounts of textual data.

2.2 U.S. Circuit Court Decisions

“Jargon serves lawyers as a bond of union: it serves them, at every word, to remind them of that common interest, by which they are made friends to one another, enemies to the rest of mankind.”

Jeremy Bentham [2, 292]

Figure 2: US Court Circuits are multi-state regions. The words of the iconoclast Bentham were not the last written on the topic of legal language. Law and language meet in many academic ways: forensic linguists help solve crimes, judges make semantic rulings on unclear contract wording, and social scholars take a high-level view, studying the language of lawyers and judges [29]. By analyzing the written decisions of the US Circuit Courts of Appeal, we hope to shed light on thematic and potentially linguistic differences between subsets of the data. Differences in word usage between courts has been previously studied using legal databases as a source for historical lexicography [8]. However, in that work, text-based searches provided information on particular words of interest. Through text mining and visualization, we select words of interest and provide a broad overview as an entry point to deeper analysis. The US Circuit Courts of Appeal are made up of 12 regionally-based court divisions (numbered First through Eleventh, plus the DC Circuit) and the Federal Circuit, which hears cases of national relevance, such as patent-related appeals (see Fig. 2). This data contains of 628,000 court decisions, each labeled by circuit. The judgments are faceted, because they can be organized along several dimensions, such as the lead authoring judge, the decision length, the date of the decision, or whether the lower court was upheld or overturned. For
Jigsaw (Multiple Documents)

Visual Analytics Support for Intelligence Analysis
Case Study: The 9/11 Report

Carsten Görg
Youn-ah Kang
Zhicheng Liu
John Stasko

Information Interfaces Group
Georgia Institute of Technology

[http://www.cc.gatech.edu/gvu/ii/jigsaw/]
Visual Analytics Support for Intelligence Analysis
Case Study: The 9/11 Report

Carsten Görg
Youn-ah Kang
Zhicheng Liu
John Stasko

Information Interfaces Group
Georgia Institute of Technology

[http://www.cc.gatech.edu/gvu/ii/jigsaw/]