Example Tasks & Goals

• Which documents contain text on topic XYZ?
• Which documents are of interest to me?
• Are there other documents that are similar to this one (so they are worthwhile)?
• How are different words used in a document or a document collection?
• What are the main themes and ideas in a document or a collection?
• Which documents have an angry tone?
• How are certain words or themes distributed through a document?
• Identify “hidden” messages or stories in this document collection.
• How does one set of documents differ from another set?
• Quickly gain an understanding of a document or collection in order to subsequently do XYZ.
• Find connections between documents.
This Week’s Agenda

Visualization for IR
Helping search

Visualizing text
Showing words, phrases, and sentences

Visualizing document sets
Words & sentences
Analysis metrics
Concepts & themes

Last Time

Related Topic - Sensemaking

• Sensemaking
  – Gaining a better understanding of the facts at hand in order to take some next steps
  – (Better definitions in VA lecture)

• InfoVis can help make a large document collection more understandable more rapidly
Today’s Agenda

- Move to collections of documents
  - Still do words, phrases, sentences
  - Add
    - More context of documents
    - Document analysis metrics
    - Document meta-data
    - Document entities
    - Connections between documents
    - Documents concepts and themes

Various Document Metrics

- Goals?
- Different variables for literary analysis
  - Average word length
  - Syllables per word
  - Average sentence length
  - Percentage of nouns, verbs, adjectives
  - Frequencies of specific words
  - Hapax Legomena – number of words that occur once

Keim & Oelke
VAST '07
Vis

Each block represents a contiguous set of words, eg, 10,000 words

Do partial overlap in blocks for a smoother appearance

The Bible

Figure 4: Visual Fingerprint of the Bible. Each pixel represents one category of the codebook of a map to the average vector length. Interesting characteristics such as the generally shorter verses of the pasture words, the long ones in the 7th books of Corinthians or the difference between the Old Testament and the New Testament can be present.

Figure 5: Visual Fingerprint of the Bible. More detailed view on the bible in which each pixel represents a single verse and verses are grouped in diagrams. What is again mapped to vector length. The detailed view reveals some interesting patterns that are not visible in the averaged version of it.
Follow-On Work

• Focus on readability metrics of documents
• Multiple measures of readability
  – Provide quantitative measures
• Features used:
  – Word length
  – Vocabulary complexity
  – Nominal forms
  – Sentence length
  – Sentence structure complexity

Visualization & Metrics

The intention of TidBits [9] is to provide a compact but yet meaningful representation of Information Retrieval results, whereas the FeatureVis technique, presented in [9], was designed to explain interesting text patterns which are suggested by the system, find meaningful co-occurrences of them, and identify their temporal evolution.

This includes aspects like ensuring contextual consistency, avoiding unknown vocabulary and difficult grammatical structures.

Figure 5: Two example sentences whose overall readability score is about the same. The detail view reveals the different reasons why the sentences are difficult to read.

Uses heatmap style vis (blue-readable, red-unreadable)
Their Paper (Before & After)

(a) The first four pages of the paper as thumbnails before the revision. (b) Detail view for one of the sections. (c) Structure thumbnails of the same pages after the revision.
Comment from the Talk

- In academic papers, you want your abstract to be really readable

- Would be cool to compare rejected papers to accepted papers

Overviews of Documents

- Can we provide a quick browsing, overview UI, maybe especially useful for small screens?
Document Cards

- Compact visual representation of a document
- Show key terms and important images

Representation

Layout algorithm searches for empty space rectangles to put things
Interaction

- Hover over non-image space shows abstract in tooltip
- Hover over image and see caption as tooltip
- Click on page number to get full page
- Click on image goes to page containing it
- Clicking on a term highlights it in overview and all tooltips
PaperLens

- Focus on academic papers
- Visualize doc metadata such as author, keywords, date, ...
- Multiple tightly-coupled views
- Analytics questions
- Effective in answering questions regarding:
  - Patterns such as frequency of authors and papers cited
  - Themes
  - Trends such as number of papers published in a topic area over time
  - Correlations between authors, topics and citations

Lee et al
CHI '05 Short
**PaperLens**

- a) Popularity of topic
- b) Selected authors
- c) Author list
- d) Degrees of separation of links
- e) Paper list
- f) Year-by-year top ten cited papers/ authors – can be sorted by topic

**NetLens**

Kang et al

*Information Visualization* ’07

Figure 1: NetLens has two symmetric windows. The left is for Content (papers) and the right for Actors (authors). Each side is further divided into panels; overview at the top, filters on the right, and lists at the bottom. Here, the Content side has two lists to reflect papers and their citations or references, and the lists on the Actor side show authors and their co-authors, respectively. The paper overview pane shows the distribution of papers (on logarithmic scales) over time, grouped by topics. Users can see which topics have the number of papers increase or decrease over 22 years. On the right side, the overview of the authors shows the distribution of countries of origin in logarithmic scale.
More Document Info

- Highlight entities within documents
  - People, places, organizations
- Document summaries
- Document similarity and clustering
- Document sentiment

Jigsaw

- Targeting sense-making scenarios
- Variety of visualizations ranging from word-specific, to entity connections, to document clusters
- Primary focus is on entity-document and entity-entity connection
- Search capability coupled with interactive exploration

Stasko, Görg, & Liu
*Information Visualization '08*
Document Cluster View

Document Grid View

Here showing sentiment analysis of docs
Calendar View

Temporal context of entities & docs

Jigsaw

- Much more to come on Visual Analytics day...
FacetAtlas

- Show entities and concepts and how they connect in a document collection
- Visualizes both local and global patterns
- Shows
  - Entities
  - Facets – classes of entities
  - Relations – connections between entities
  - Clusters – groups of similar entities in a facet

Cao et al
TVCG (InfoVis) ’10

Visualization

Video
Up to Higher Level

- How do we present the contents, semantics, themes, etc of the documents
  - Someone may not have time to read them all
  - Someone just wants to understand them

- Who cares?
  - Researchers, fraud investigators, CIA, news reporters

Vector Space Analysis

- How does one compare the similarity of two documents?
  - One model
    - Make list of each unique word in document
      - Throw out common words (a, an, the, ...)
      - Make different forms the same (bake, bakes, baked)
    - Store count of how many times each word appeared
      - Alphabetize, make into a vector
Vector Space Analysis

- Model (continued)
  - Want to see how closely two vectors go in the same direction, inner product
  - Can get similarity of each document to every other one
  - Use a mass-spring layout algorithm to position representations of each document
- Some similarities to how search engines work

Wiggle

- Not all terms or words are equally useful
- Often apply TFIDF
  - Term frequency, inverse document frequency
- Weight of a word goes up if it appears often in a document, but not often in the collection
**Process**

Documents → Analysis (Decomposition, statistics) → Algorithms (Similarity, clustering, normalization) → Visualization (Data tables for vis, 2D, 3D display)

---

**Smart System**

- Uses vector space model for documents
  - May break document into chapters and sections and deal with those as atoms
- Plot document atoms on circumference of circle
- Draw line between items if their similarity exceeds some threshold value

Salton et al
*Science '95*
Text Relation Maps

- Label on line can indicate similarity value
- Items evenly spaced
- Doesn’t give viewer idea of how big each section/document is

Improved Design

Proportional to length of section
Links placed at correct relative position
Text Themes

- Look for sets of regions in a document (or sets of documents) that all have common theme
  - Closely related to each other, but different from rest

- Need to run clustering process

Algorithm

- Recognize triangles in relation maps
  - Three with edges above threshold
- Make a new vector that is centroid of 3
- Triangles merged whenever centroid vectors are sufficiently similar
Text Theme Example

- Triangles shown
- Colored in to help presentation

Skimming and Summarization

- Can use graph traversal to follow specific themes throughout collection
- Walk along connected edges
VIBE System

- Smaller sets of documents than whole library
- Example: Set of 100 documents retrieved from a web search
- Idea is to understand contents of documents relate to each other

Focus

- Points of Interest
  - Terms or keywords that are of interest to user
    - Example: cooking, pies, apples
- Want to visualize a document collection where each document’s relation to points of interest is show
- Also visualize how documents are similar or different
**Technique**

- Represent points of interest as vertices on convex polygon
- Documents are small points inside the polygon
- How close a point is to a vertex represents how strong that term is within the document

**Algorithm**

- Example: 3 POIs
- Document \((P1, P2, P3)\) \((0.4, 0.8, 0.2)\)
- Take first two

\[
\frac{0.4}{0.4+0.8} = 0.333
\]

1/3 of way from P2 to P1
**Algorithm**

- Combine weight of first two 1.2 and make a new point, $P'$
- Do same thing for third point

\[ \frac{1.2}{1.2 + 0.2} = 0.86 \]

0.14 of way from $P'$ to $P_3$

**Sample Visualization**
VIBE Pro’s and Con’s

- Effectively communications relationships
- Straightforward methodology and vis are easy to follow
- Can show relatively large collections
- Not showing much about a document
- Single items lose “detail” in the presentation
- Starts to break down with large number of terms

Visualizing Documents

- VIBE presented documents with respect to a finite number of special terms
- How about generalizing this?
  - Show large set of documents
  - Any important terms within the set become key landmarks
  - Not restricted to convex polygon idea
Basic Idea

• Break each document into its words
• Two documents are “similar” if they share many words
• Use mass-spring graph-like algorithm for clustering similar documents together and dissimilar documents far apart

Kohonen’s Feature Maps

• AKA Self-Organizing Maps
• Expresses complex, non-linear relationships between high dimensional data items into simple geometric relationships on a 2-d display
• Uses neural network techniques

Lin
Visualization '92
Map Display of SOM

Map Attributes

- Different, colored areas correspond to different concepts in collection
- Size of area corresponds to its relative importance in set
- Neighboring regions indicate commonalities in concepts
- Dots in regions can represent documents
Work at PNNL

- Group has developed a number of visualization techniques for document collections
  - Galaxies
  - Themescapes
  - ThemeRiver
  - ...

Wise et al
InfoVis '95

Galaxies
Presentation of documents where similar ones cluster together
**Themescape**

- Self-organizing maps didn’t reflect density of regions all that well -- Can we improve?
- Use 3D representation, and have height represent density or number of documents in region
Related Topic

- Maps of Science
- Visualize the relationships of areas of science, emerging research disciplines, the impact of particular researchers or institutions, etc.
- Often use documents as the “input data”
Wonderful Book and Website

Atlas of Science
Visualizing What We Know

K. Börner

http://scimaps.org

Some Examples

Boyack & Klavans

http://scimaps.org/maps/map/map_of_scientific_pa_55/
Science Related Wikipedia Activity

http://scimaps.org/maps/map/science_related_wiki_49/

http://scimaps.org/maps/map/maps_of_science_fore_50/
Temporal Issues

- Semantic map gives no indication of the chronology of documents
- Can we show themes and how they rise or fall over time?

---

ThemeRiver

Havre, Hetzler, & Nowell
InfoVis '00
Representation

- Time flows from left->right
- Each band/current is a topic or theme
- Width of band is “strength” of that topic in documents at that time

More Information

- What’s in the bands?
- Analysts may want to know about what each band is about
TIARA

- Keeps basic ThemeRiver metaphor
- Embed word clouds into bands to tell more about what is in each
- Magnifier lens for getting more details
- Uses Latent Dirichlet Allocation to do text analysis and summarization
Features

Lens shows email senders & receivers

Documents containing “cetable”

Reminders

• HW 5 due next Thursday
  - Questions?

• Project midway report due in 2 weeks
Upcoming

• Interaction 1
  – Reading
    Few chapter 4
    Yi et al ’07

• Interaction 2