Text and Document Visualization 2

CS 4460 – Intro. to Information Visualization
November 15, 2017
John Stasko

Learning Objectives

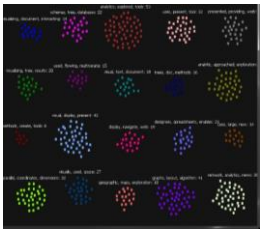
- Explain what word concordance is & how WordTree representation works
- List different queries/tasks often needed on document collections
- List various analytic metrics often calculated on documents
- List different aspects of documents often visualized
- Explain vector space document analysis (similarity calculation, search) & TFIDF
- Describe visual representation used by and contributions of these systems
  - SentenTree, TextArc, Themail, Jigsaw, ThemeScape/IN-SPIRE, ThemeRiver

This Week's Agenda

Visualizing text
Showing words, combinations, and context

→

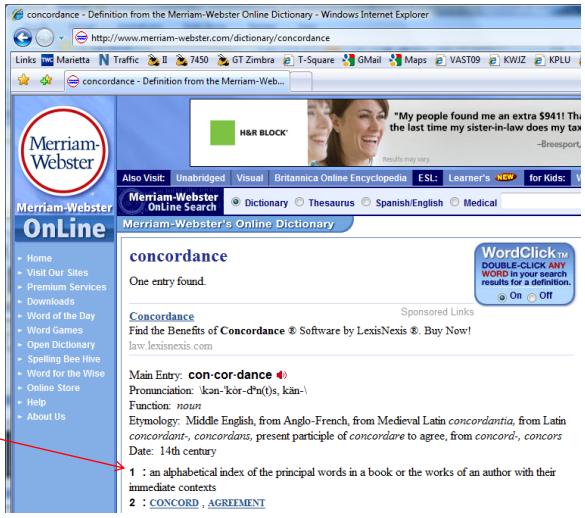
Visualizing document sets
Words & sentences
Analysis metrics
Concepts & themes



Beyond Individual Words

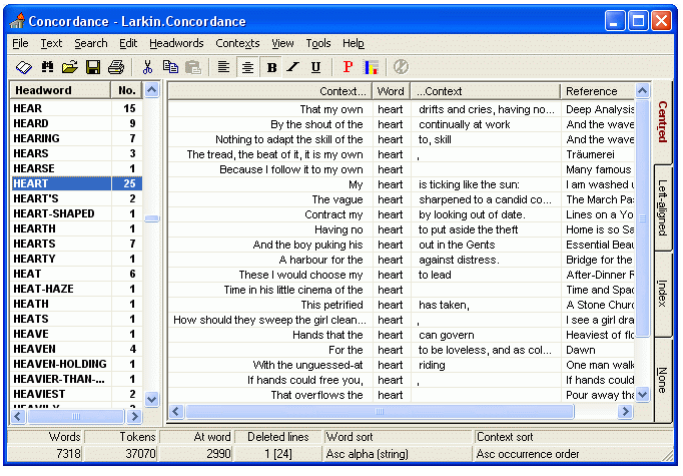
- The techniques (from last class) focus largely on words
  - Especially word clouds & wordles
- Can we show combinations of words, ie, actual phrases and sentences, in order to provide more context?

Concordance



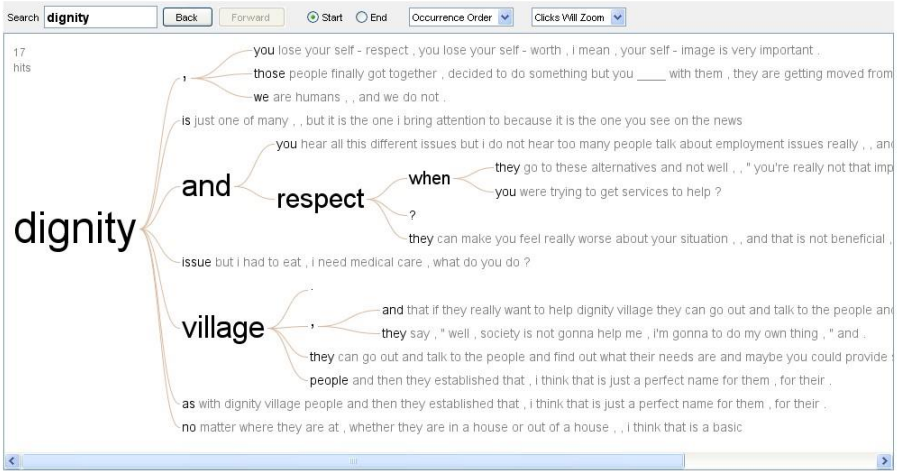
Definition

Concordance in Text



http://www.concordancesoftware.co.uk

Word Tree

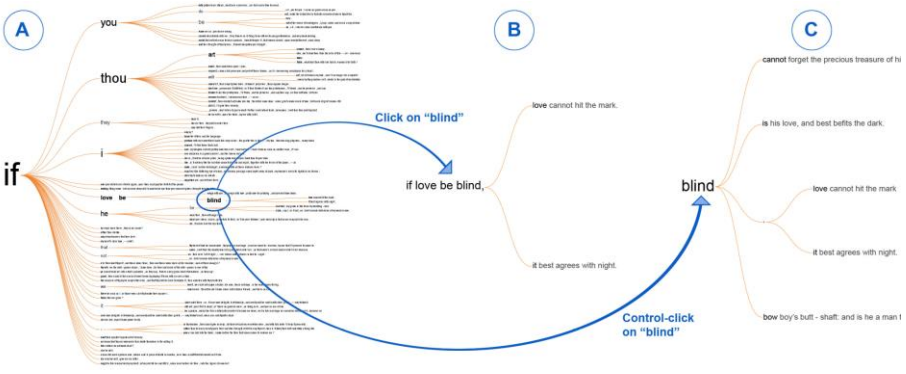


Word Tree

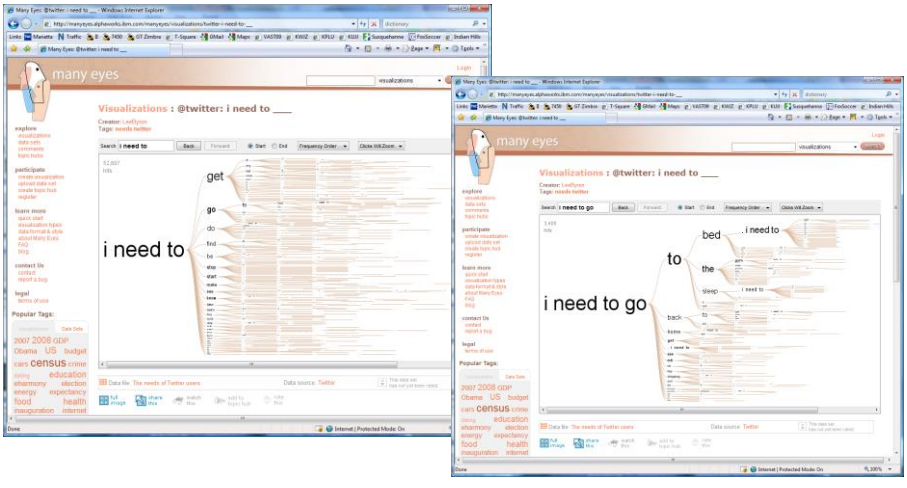
- Shows context of a word or words
  - Follow word with all the phrases that follow it
- Font size shows frequency of appearance
- Continue branch until hitting unique phrase
- Clicking on phrase makes it the focus
- Ordered alphabetically, by frequency, or by first appearance

Wattenberg & Viégas
TVCG (InfoVis) '08

Interaction



Many Eyes' WordTree



Words and Context

- Can we show most frequent words like a word cloud but also provide context?
  - Should each word appear one time?
  - But then how to show context?
  - If appears multiple times, how to make that work?

SentenTree

Hu, Wongsuphasawat, and Stasko
TVCG '17 (InfoVis '16)

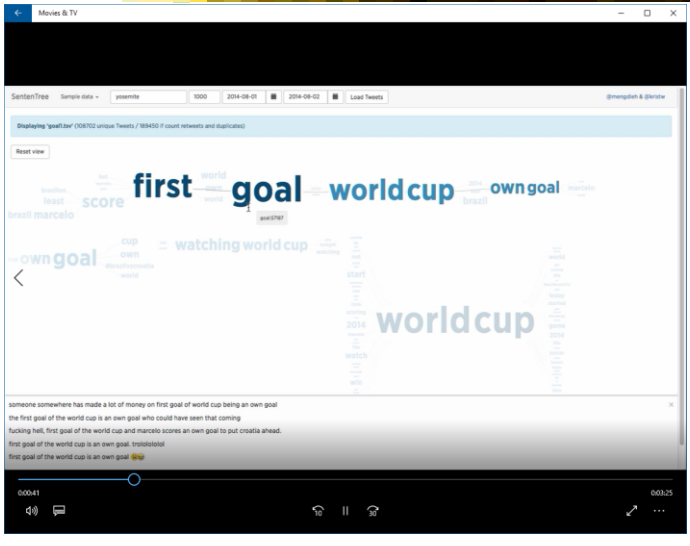
- Elements of word clouds and word trees
  - Highlight keywords using size
  - Show sentence fragments
  - Provide a summary of the dataset
  - Enable interactive drill-down into details

Summary of 189,450 tweets (108,702 unique) posted in a 15 minute time window around the first goal of the opening game of the 2014 Soccer World Cup

Example

Tweets mentioning word "Yosemite" from Aug 1, 2014

Video

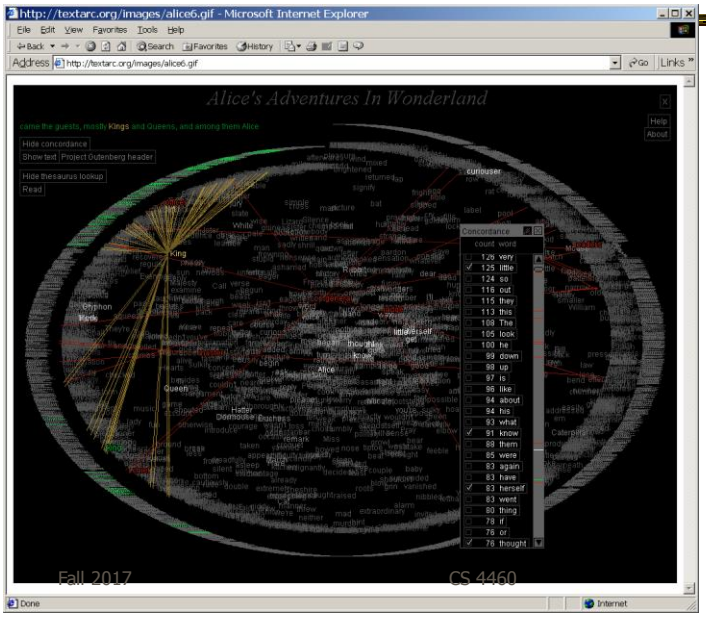


Another Challenge

- Visualize an entire book
- What does that mean?
  - Word appearances
  - Sentences
  - …

TextArc

http://textarc.org



Sentences laid out in order of appearance

Words near to where they appear

Significant interaction

Brad Paley

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

Exercise

- Suppose you have 300 reviews of a TV/LCD monitor on Amazon?
  - Want to build a visualization of these…but
- What tasks/goals/questions would someone want to accomplish with the vis?
- Let's generate a list…

Questions/Tasks

What is the tone of a review? (sentiment analysis)
What words show up in the reviews, especially in pertiment subsets?
What was avg score?
Facilitate comparisons.
Find features, see what words go with each.
Distribution of scores.
How have scores changed over time?
Surface pictures of product.
Show attributes/metrics of diff reviews
Represent writer's background/history.
If available, compare/contrast public vs expert views
Check out certain important words "recommend", "suggest"
Look for frequencies of certain key words.
Make actual text available.

Overlaps & Similarities

- Are some items in our list in the same "category"?
  - Can we generalize a little and narrow the list down to some core questions/tasks?

Evaluate a Vis

- Use the generated list to evaluate a visualization for this problem
- Think about this for your P5…

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.
- Understand the history of changes in a document.
- Find connections between documents.

Various Document Metrics

- 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

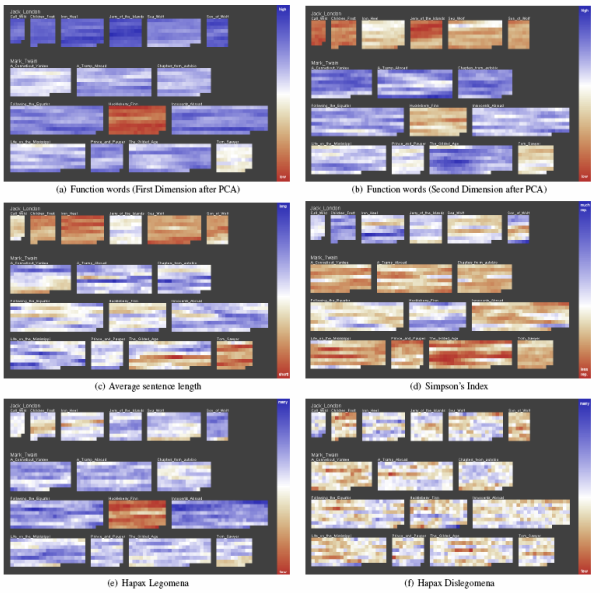
How would you visualize this?

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



(a) Function words (First Dimension after PCA)
(b) Function words (Second Dimension after PCA)
(c) Average sentence length
(d) Simpson's Index
(e) Hapax Legomena
(f) Hapax Dislegomena

Figure 2: Fingerprints of books of Mark Twain and Jack London. Different measures for authorship attribution are tested. If a measure is able to discriminate between the two authors, the visualizations of the books that are written by the same author will equal each other more than the visualizations of books written by different authors. It can easily be seen that this is not true for every measure (e.g. Hapax Dislegomena). Furthermore, it is interesting to observe that the book *Huckleberry Finn* sticks out in a number of measures as if it is not written by Mark Twain.

The Bible

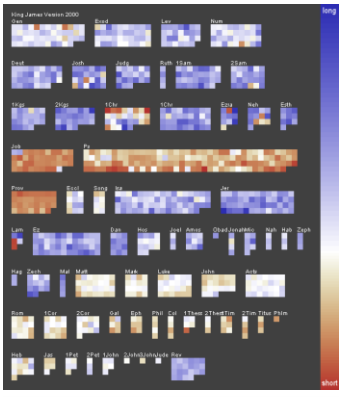


Figure 4: Visual Fingerprint of the Bible. Each pixel represents one chapter of the bible and color is mapped to the average verse length. Interesting characteristics such as the generally shorther verses of the poetry books, the inhomogeneity of the 1. Book of Chronicles or the difference between the Old Testament and the New Testament can be perceived.

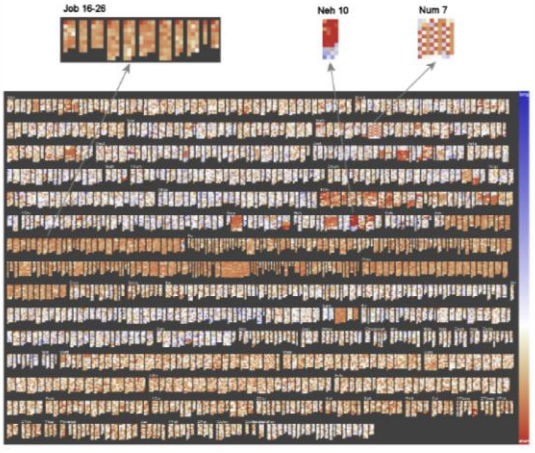
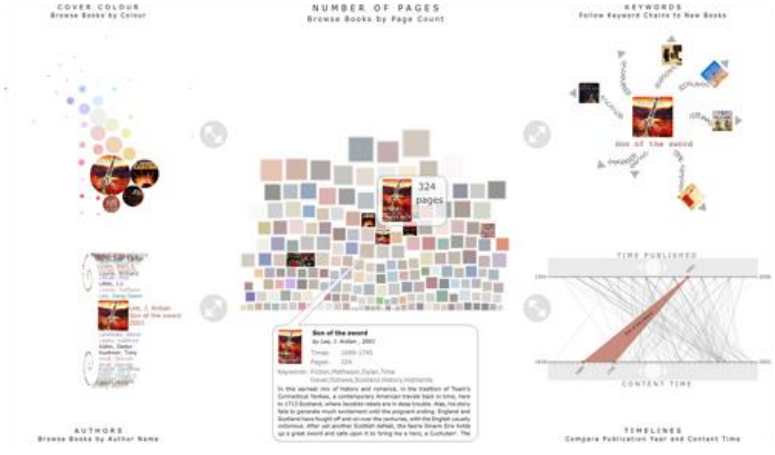


Figure 5: Visual Fingerprint of the Bible. More detailed view on the bible in which each pixel represents a single verse and verses are grouped to chapters. Color is again mapped to verse length. The detailed view reveals some interesting patterns that are camouflaged in the averaged version of fig. 4.

Bohemian Bookshelf

Video

Serendipitous browsing



Thudt et al
CHI '12

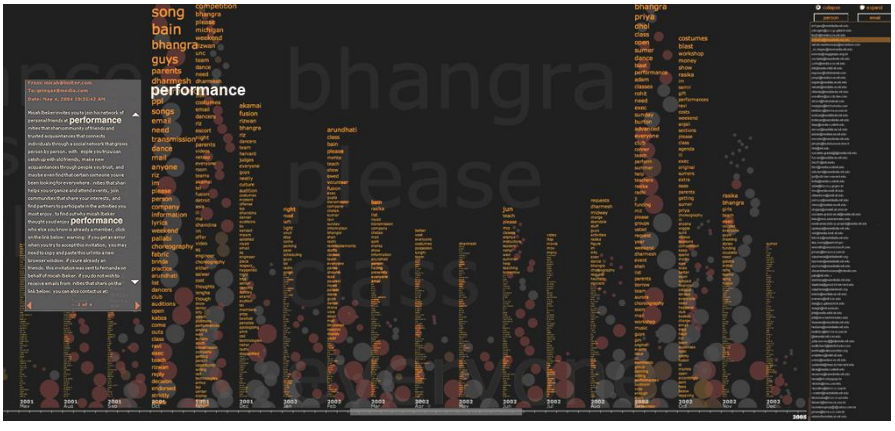
Themail

- Visualize one's email history
  - With whom and when has a person corresponded
  - What words were used
- Answer questions like:
  - What sorts of things do I (the owner of the archive) talk about with each of my email contacts?
  - How do my email conversations with one person differ from those with other people?

Viégas, Golder & Donath
CHI '06

Interface

Text analysis to seed visualization
Monthly & yearly words



More Document Info

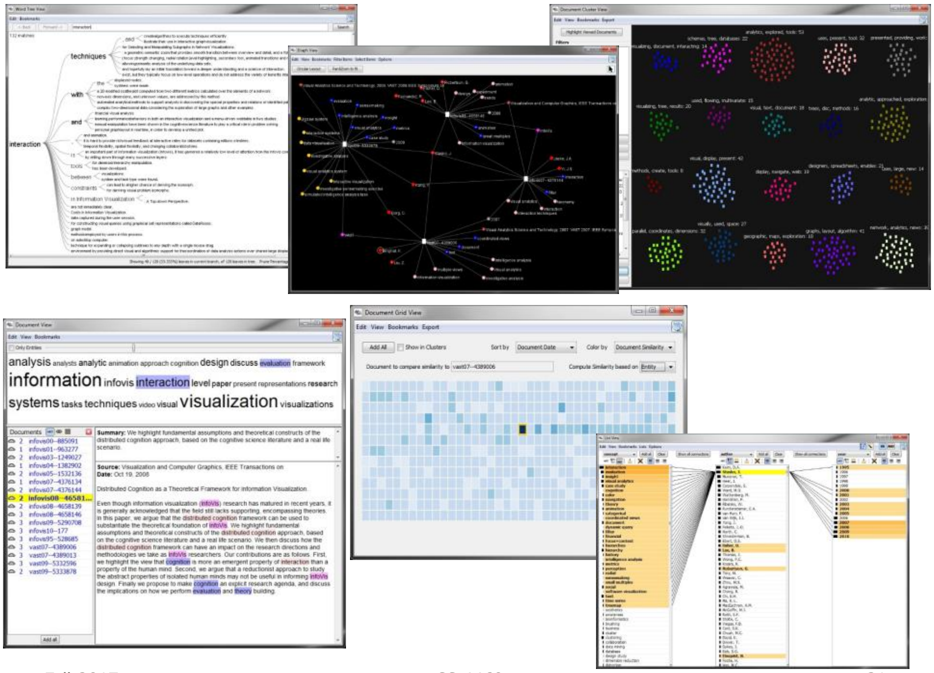
- Highlight entities within documents
  - People, places, organizations
- Document summaries
- Document similarity and clustering
- Document sentiment

Jigsaw

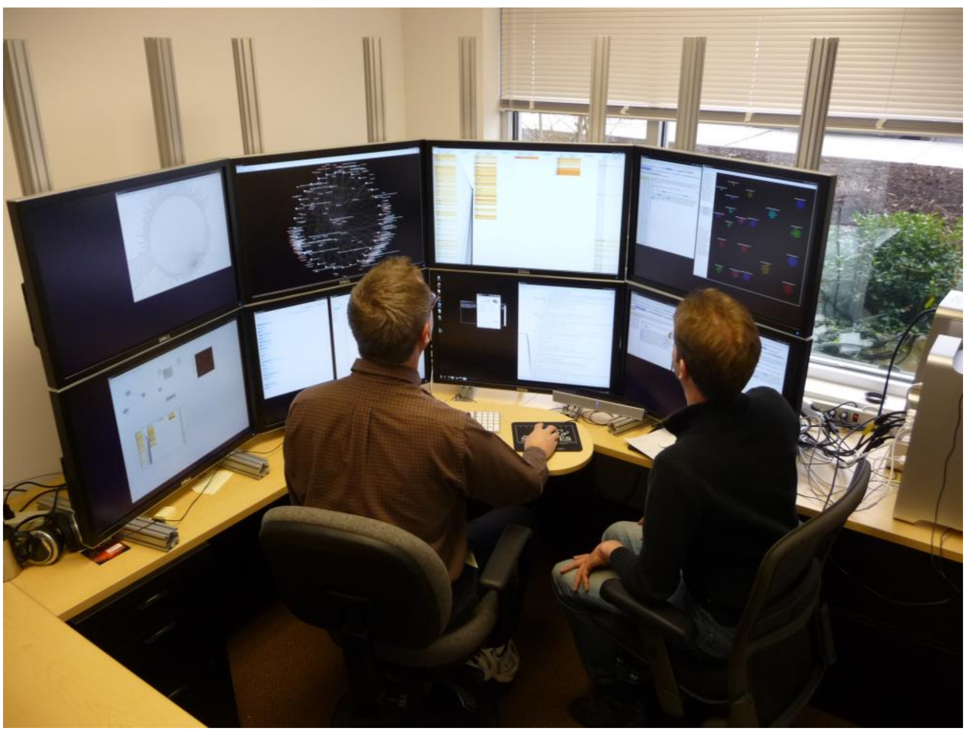
Saw earlier in term

- 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
Görg et al
IEEE TVCG '13



Document Grid View
analysis analysts analytic animation approach cognition design discuss evaluation framework
information infovis interaction level paper present representations research
systems tasks techniques video visual visualization visualizations

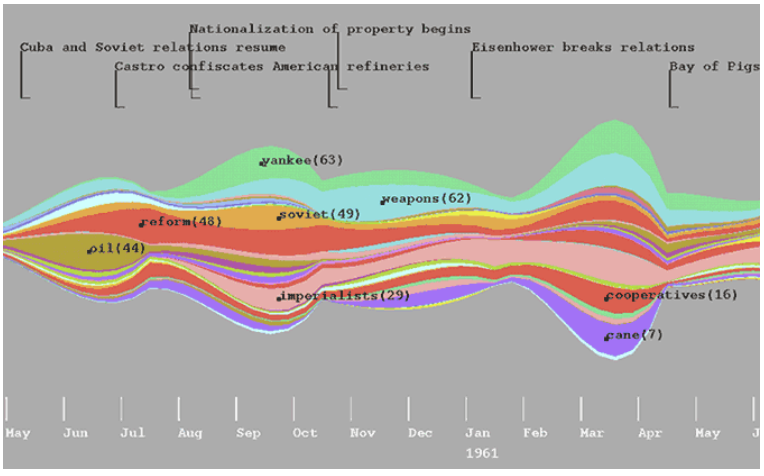


Temporal Issues

- What if documents have dates?
- Can we visualize their contents over time?

ThemeRiver

Each stream is a word or a theme



Havre, Hetzler, & Nowell InfoVis '00

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 same direction, inner product
  - Can get similarity of each document to every other one
- 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

Document Collection Maps

- Use document similarity idea
- Use mass-spring graph-like algorithm for clustering similar documents together and moving dissimilar documents far apart

Work at PNNL

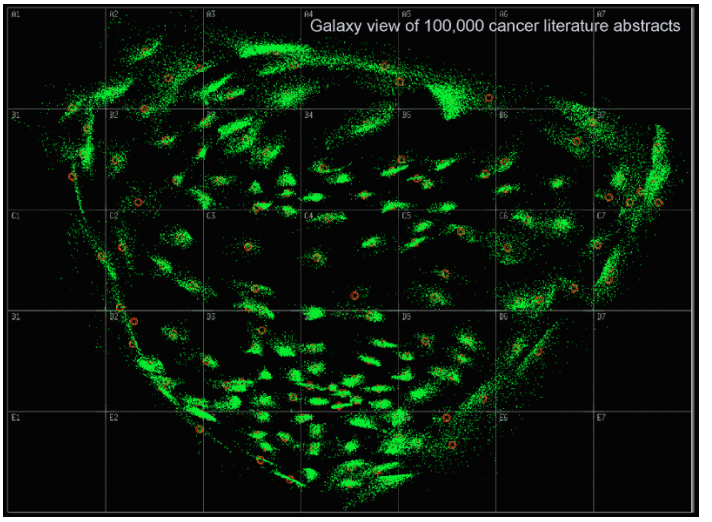
http://www.pnl.gov/infoviz

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

Wise et al
InfoVis '95

Galaxies

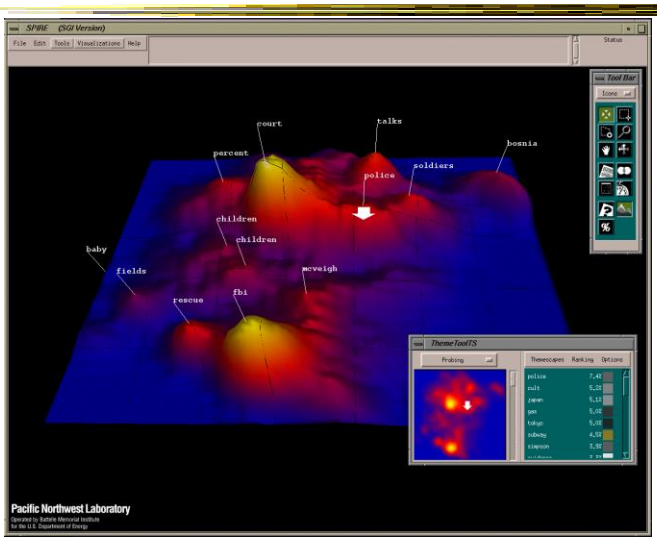
Presentation of documents where similar ones cluster together



Themescapes

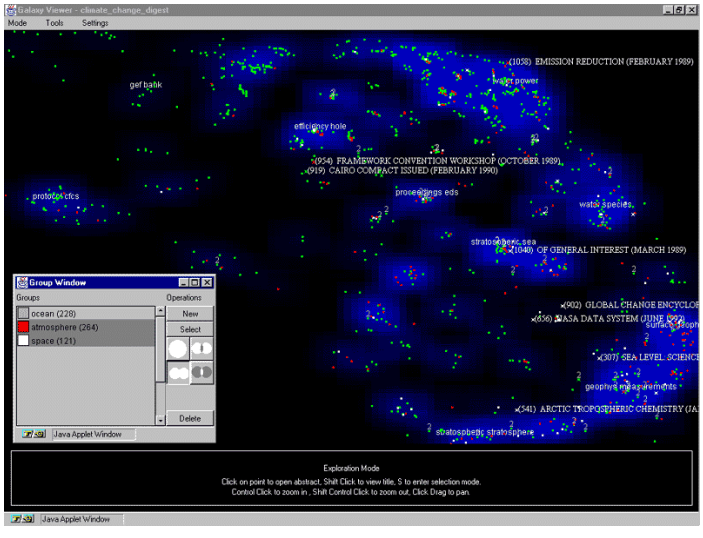
- Self-organizing maps didn't reflect density of regions all that well -- Can we improve?
- Use 3D representation, and have height represent density or number of documents in region

Themescape



Video

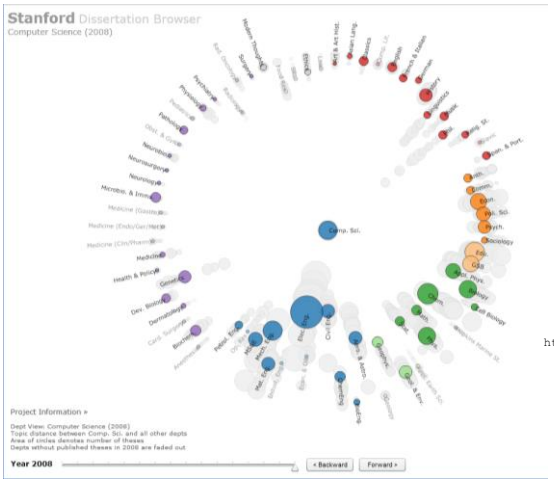
WebTheme



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"

Stanford Diss. Browser



9,000 Stanford PhD theses

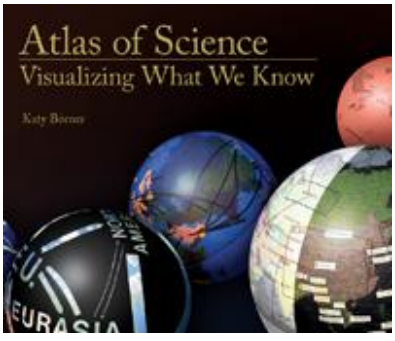
Rather than overall semantic map, you choose a focus and all update to show their relationship

Demo at
http://nlp.stanford.edu/projects/dissertations/

Chuang et al
CHI '12

http://nlp.stanford.edu/projects/dissertations/browser.html

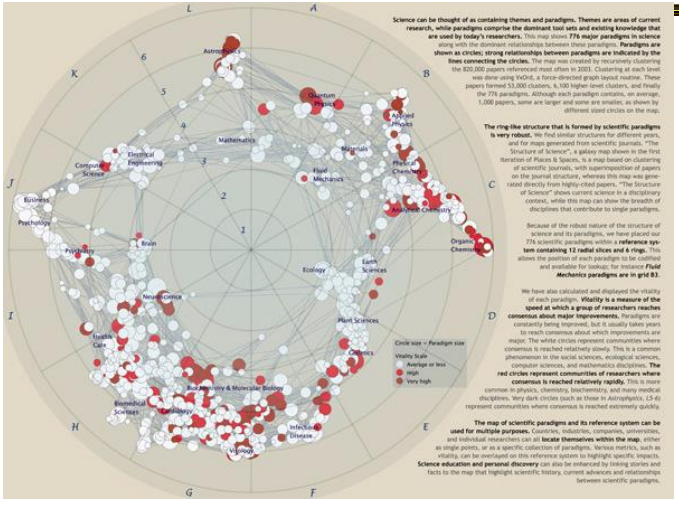
Wonderful Book and Website



K. Börner

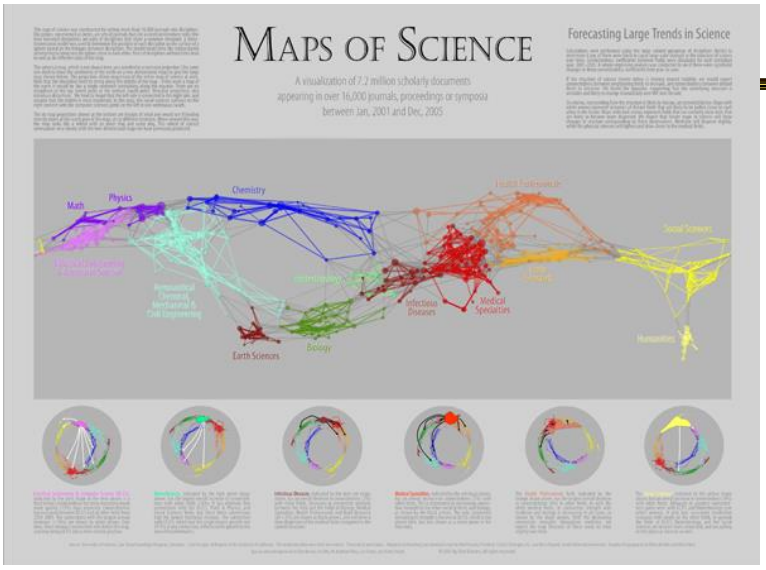
http://scimaps.org

Some Examples



Boyack & Klavans

http://scimaps.org/maps/map/map_of_scientific_pa_55/

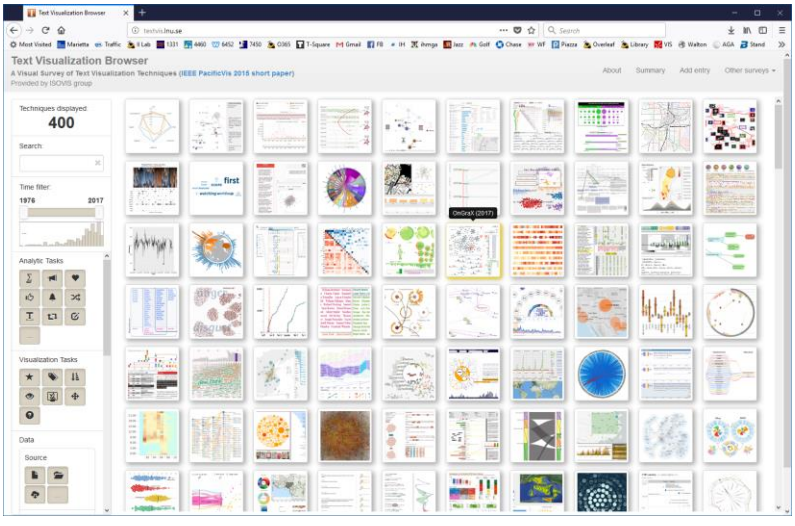


Klavans & Boyack

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

http://textvis.lnu.se

Text Vis Browser



Learning Objectives

- Explain what word concordance is & how WordTree representation works
- List different queries/tasks often needed on document collections
- List various analytic metrics often calculated on documents
- List different aspects of documents often visualized
- Explain vector space document analysis (similarity calculation, search) & TFIDF
- Describe visual representation used by and contributions of these systems
  - SentenTree, TextArc, Themail, Jigsaw, ThemeScape/IN-SPIRE, ThemeRiver

Programs

- Some thoughts…

Upcoming

- Lab 9 – Layout in D3 (networks)
  - Prep: Force-directed Graphs website
- Casual InfoVis
  - Prep: Watch InfoCanvas video