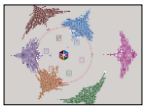
Text and Document Visualization 2

CS 7450 - Information Visualization
November 2, 2016
John Stasko

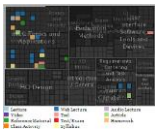
Learning Objectives

- 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)
- Explain TFIDF
- Describe visual representation used by and contributions of these systems
  - Themail, PaperLens, Jigsaw, ThemeScape/IN-SPIRE, ThemeRiver

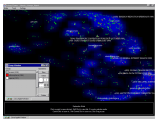
This Week's Agenda



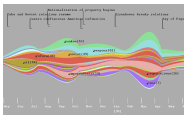
Visualization for IR
Helping search



Visualizing text
Showing words, phrases, and sentences



Visualizing document sets
Words & sentences
Analysis metrics
Concepts & themes



Last Time

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

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

Your HW

- What tasks/goals/questions could be expected with the Amazon reviews data set?
- Let's generate a list…

Questions/Tasks

How many reviewers recommend the TV?
How do the reviewers rate the different features of the TV?
How long has a reviewer had the TV?
How many five star reviews are there for the TV?
What are the favorite and least favorite factors for the TV?
Was there anything associated with the word 'annoying'?
How long did it take to get a negative review?
Do they recommend it to others?
How does it compare to other TVs?
What do users think about the size?
Is it good value for money?
For bad reviews, what are the frequent negative reviews?
Do users have a particular use for the TV?
If it has problems, are they easy to fix?
Was delivery quick and easy?
How many reviews speak about a specific feature?
How was the setup process?
How many reviews talk about the purchase and service experience?
What is the average rating?
Is it a good TV or not?
How is the feature X rated?
Have the review sentiments changed over time?
What do they think of the brand / compare it to others?

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?

Questions/Tasks

Is this a good TV? (summarize ratings)
What are the positive/negative features?
Characterize product features vs other stuff.
What do people think about X?
Comparison to other TVs.
What are the defects/problems?
What are the views over time?

Evaluate a Vis

- Use that list to evaluate a visualization for this problem

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

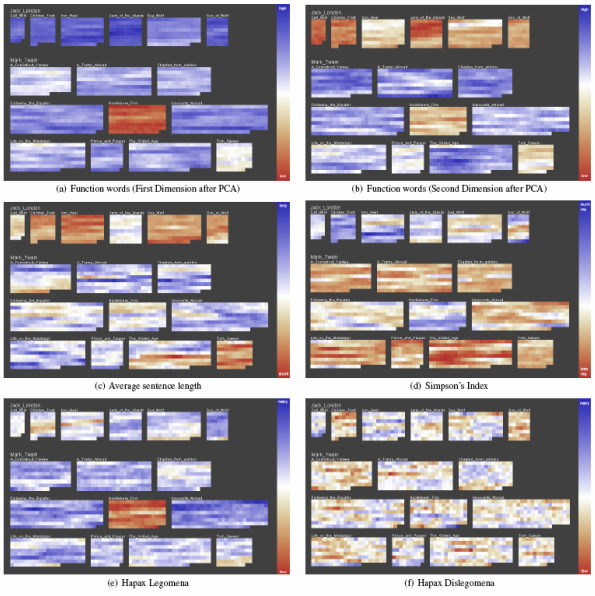
- Different variables for literary analysis
  - Average word length
  - Syllables per word
  - Average sentence length
  - Percentage of nouns, verbs, adjectives
  - Frequencies of specific words
  - Hapax Legomena – number of words that occur once

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

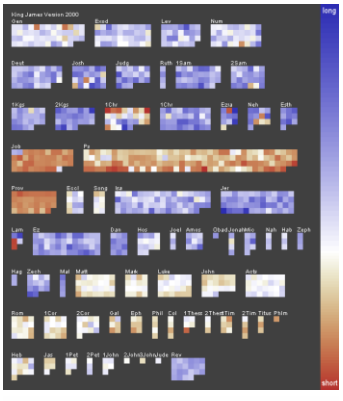


Figure 4: Visual Fingerprint of the Bible. Each pixel represents one chapter of the bible and color is mapped to the average verse length. Interesting characteristics such as the generally shorther verses of the poetry books, the inhomogeneity of the 1. Book of Chronicles or the difference between the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

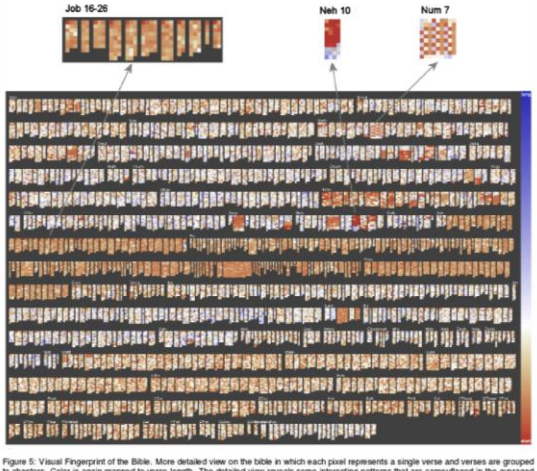


Figure 5: Visual Fingerprint of the Bible. More detailed view on the bible in which each pixel represents a single verse and verses are grouped to chapters. Color is again mapped to verse length. The detailed view reveals some interesting patterns that are camouflaged in the averaged version of fig.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

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

Oelke & Keim
VAST '10

Visualization & Metrics

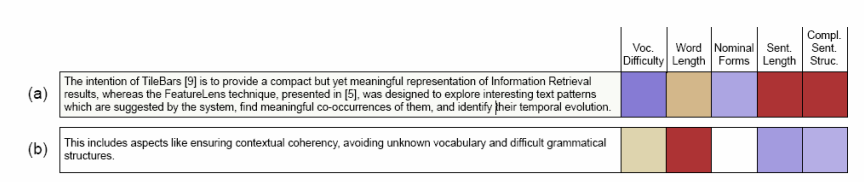


Figure 5: Two example sentences whose overall readability score is about the same. The detail view reveals the different reasons why the sentences are difficult to read.

Uses heatmap style vis (blue-readable, red-unreadable)

Interface

Figure 3: Screenshot of the VisRA tool on 3 different aggregation levels. (a) Corpus View (b) Block View (c) Detail View. To display single features, the colormap is generated as described in section 3.4 and figure 2.

Their Paper (Before & After)

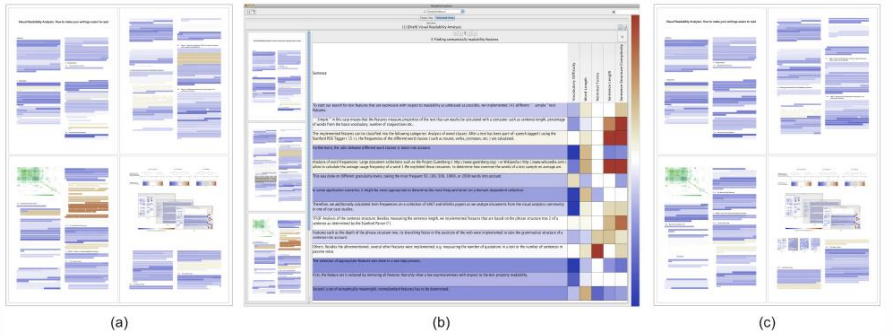


Figure 6: Revision of our own paper. (a) The first four pages of the paper as structure thumbnails before the revision. (b) Detail view for one of the sections. (c) Structure thumbnails of the same pages after the revision.

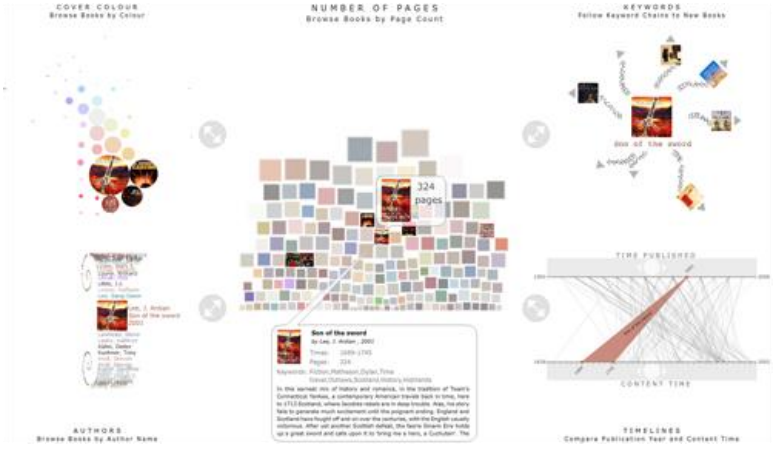
Comment from the Talk

- In academic papers, you want your abstract to be really readable
- Would be cool to compare rejected papers to accepted papers

Bohemian Bookshelf

Video

Serendipitous browsing



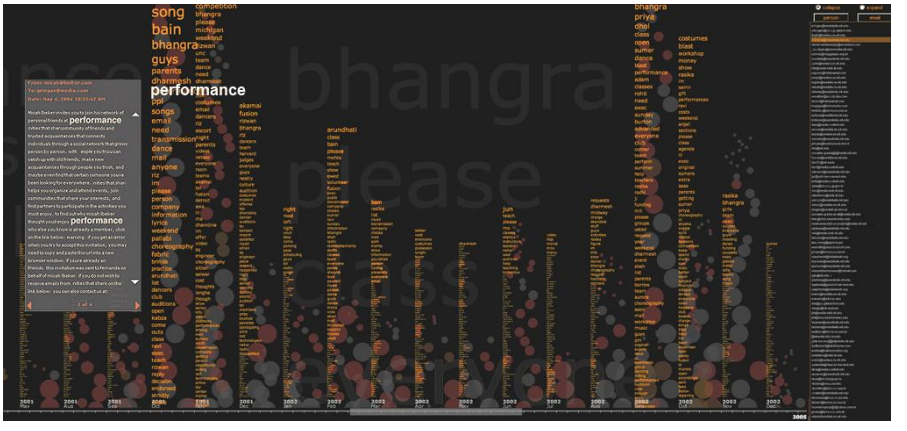
Thudt et al
CHI '12

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



Characteristics

- Text analysis to seed visualization
- Monthly & yearly words

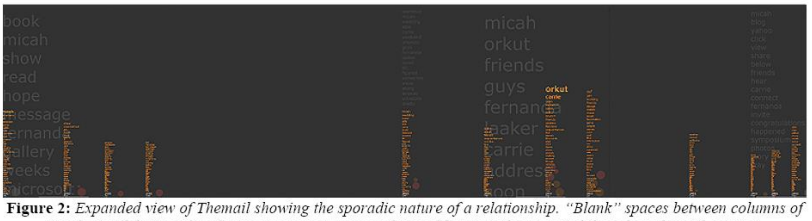


Figure 2: *Expanded view of Themail showing the sporadic nature of a relationship. "Blank" spaces between columns of words stand for months when no messages were exchanged between the user and the selected email contact.*

Query UI

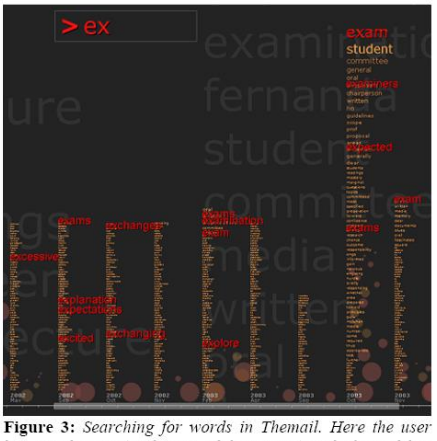


Figure 3: *Searching for words in Themail. Here the user has typed "ex" (at the top of the screen) and Themail has highlighted (in red) all the monthly words starting with these characters.*

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

Video

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

More Document Info

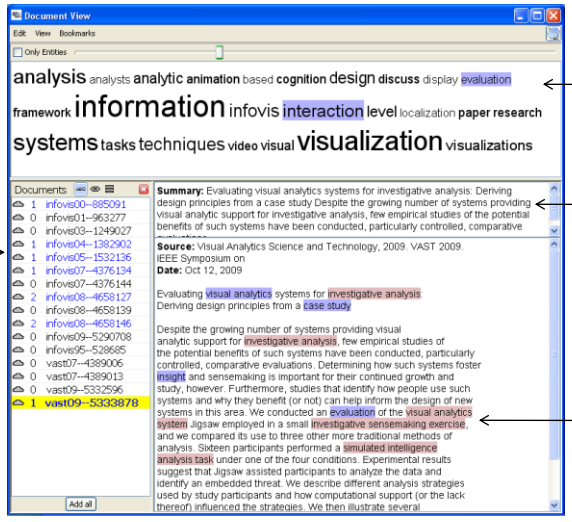
- Highlight entities within documents
  - People, places, organizations
- Document summaries
- Document similarity and clustering
- Document sentiment

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
Görg et al
IEEE TVCG '13

Document View



Wordcloud overview

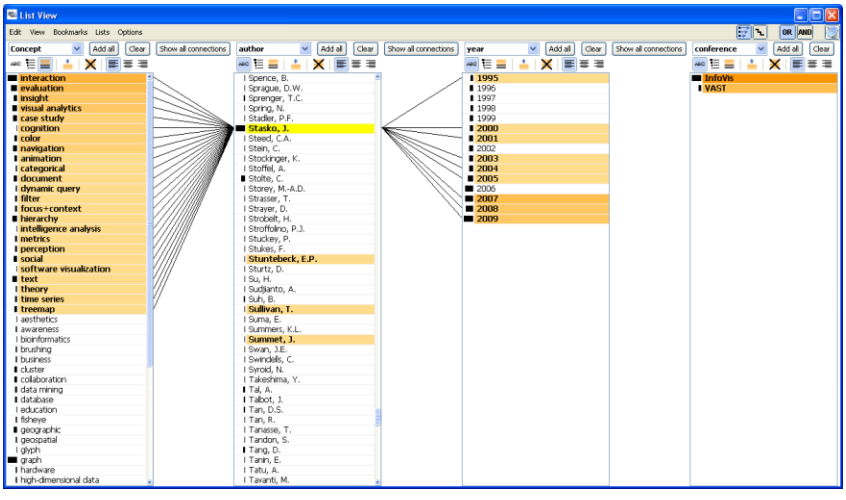
Doc List

Document summary

Selected document's text with entities identified

List View

Entities listed by type



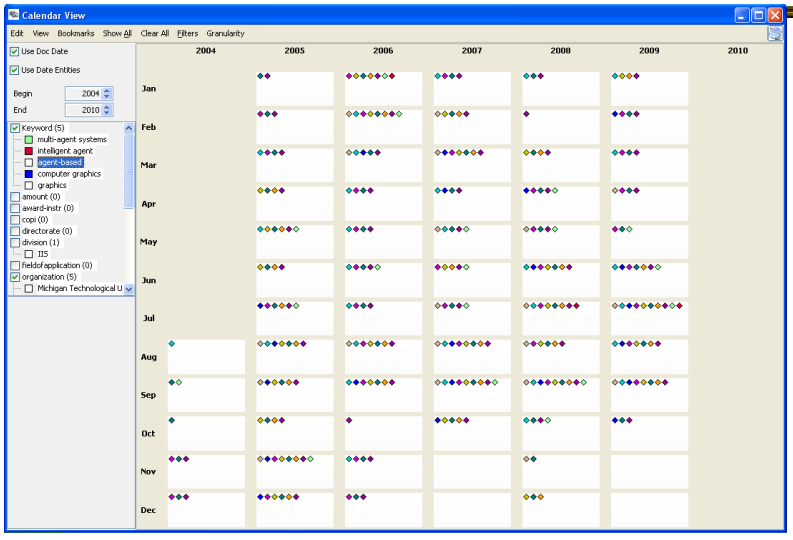
Document Cluster View

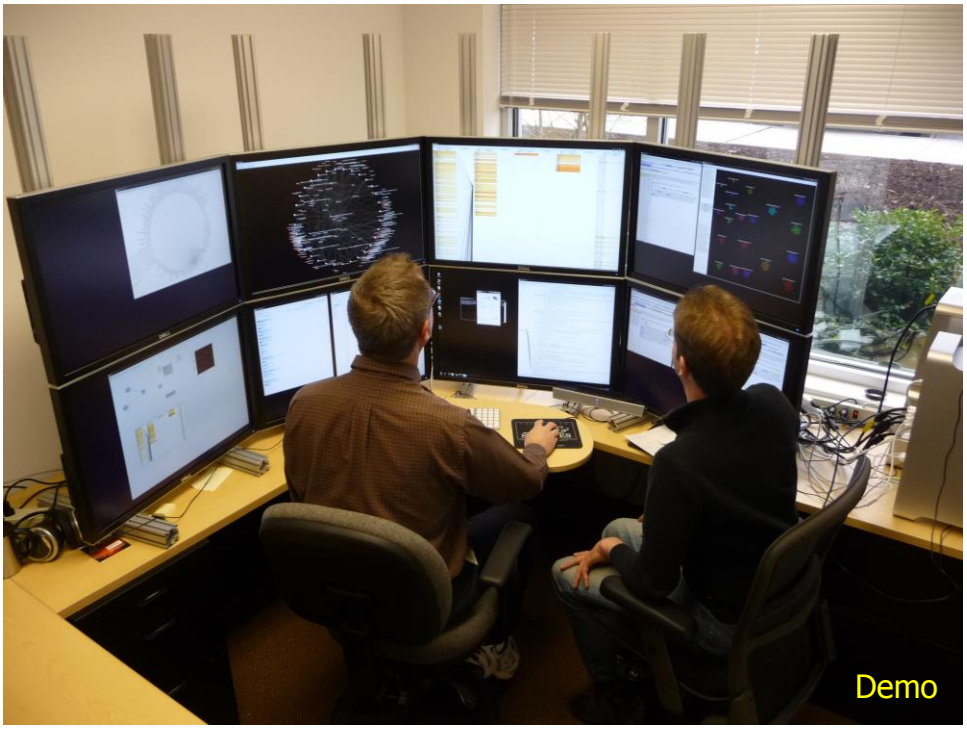
Document Grid View

Here showing sentiment analysis of docs

Calendar View

Temporal context
of entities & docs





Jigsaw

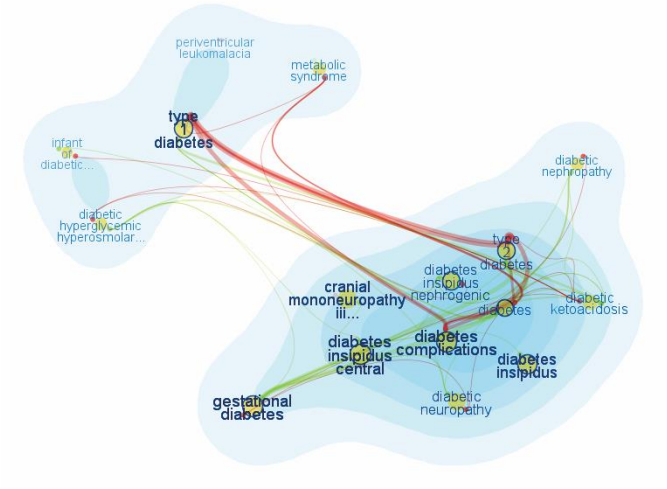
- More to come on Visual Analytics day…

FacetAtlas

- Show entities and concepts and how they connect in a document collection
- Visualizes both local and global patters
- 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



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

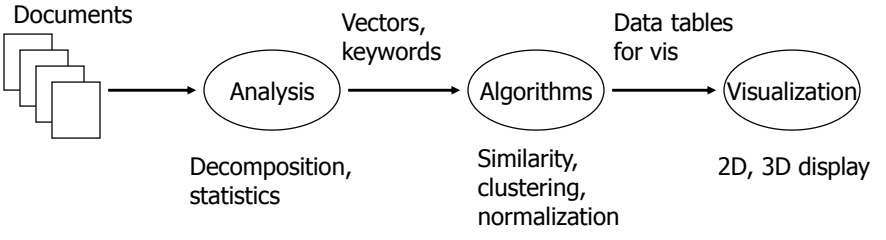
Vector Space Analysis

- 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
  - 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



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

Olsen et al
Info Process & Mgmt '93

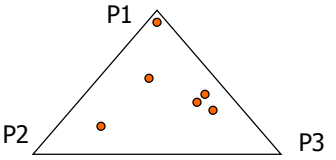
Focus

- 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

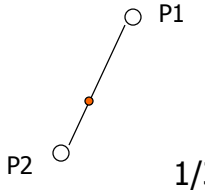
Technique

- 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

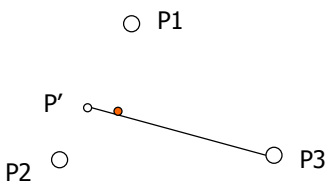


$$\frac{0.4}{0.4+0.8} = 0.333$$

1/3 of way from P2 to P1

Algorithm

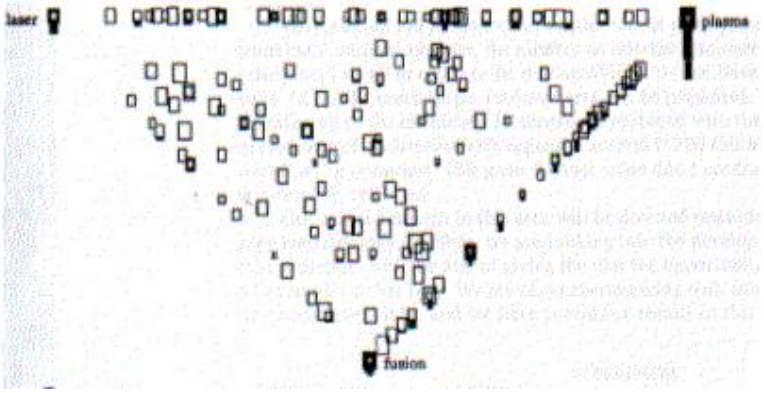
- Combine weight of first two 1.2 and make a new point, P'
- Do same thing for third point



$$\frac{1.2}{1.2+0.2} = 0.86$$

0.14 of way from P' to P3

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

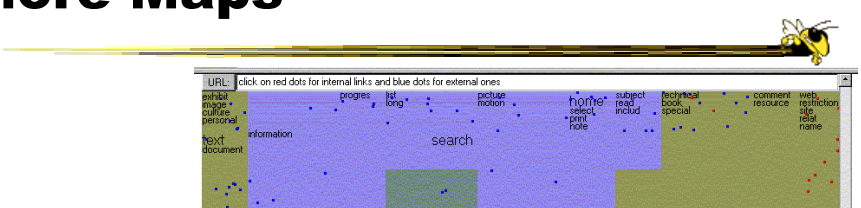
- AKA Self-Organizing Maps
- Expresses complex, non-linear relationships between high dimensional data items into simple geometric relationships on a 2-d display
- Uses neural network techniques

Lin
Visualization '92

Map 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

More Maps



Interactive demos

Xia Lin

Work at PNNL

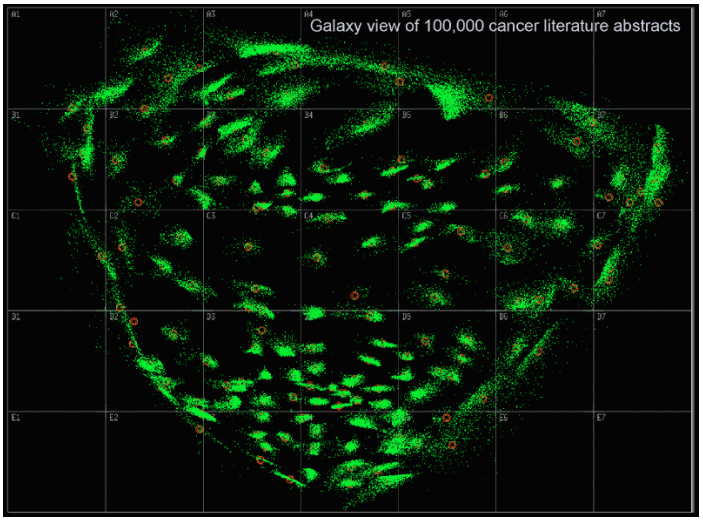
http://www.pnl.gov/infoviz

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

Wise et al
InfoVis '95

Galaxies

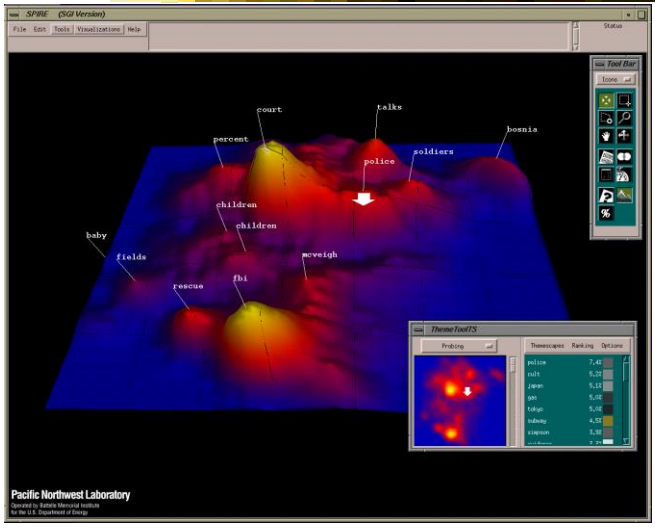
Presentation of documents where similar ones cluster together



Themescapes

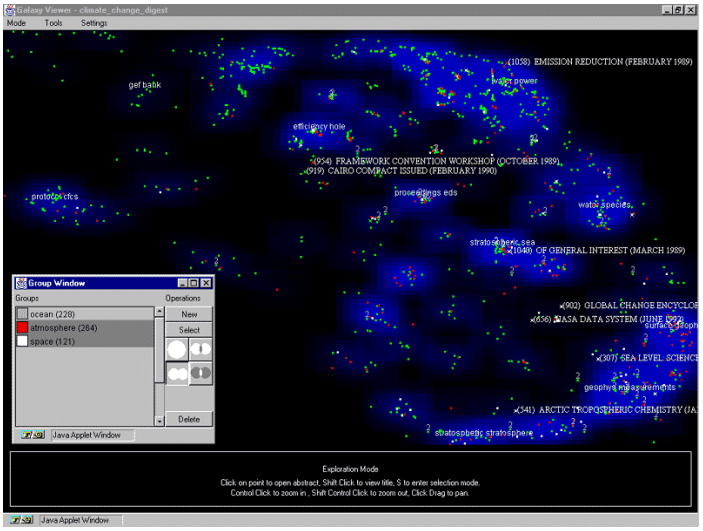
- Self-organizing maps didn't reflect density of regions all that well -- Can we improve?
- Use 3D representation, and have height represent density or number of documents in region

Themescape



Video

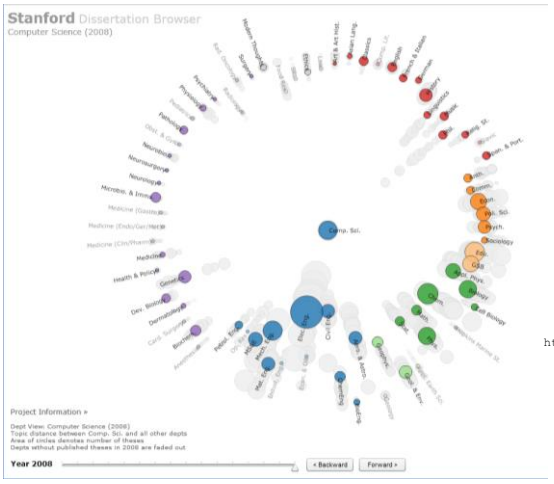
WebTheme



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

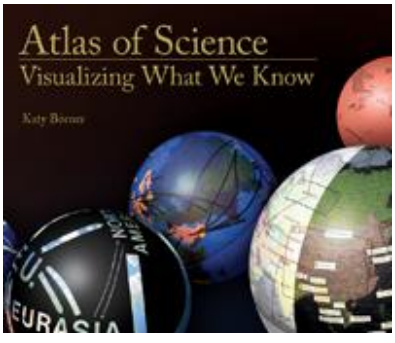
Rather than overall semantic map, you choose a focus and all update to show their relationship

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

Chuang et al
CHI '12

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

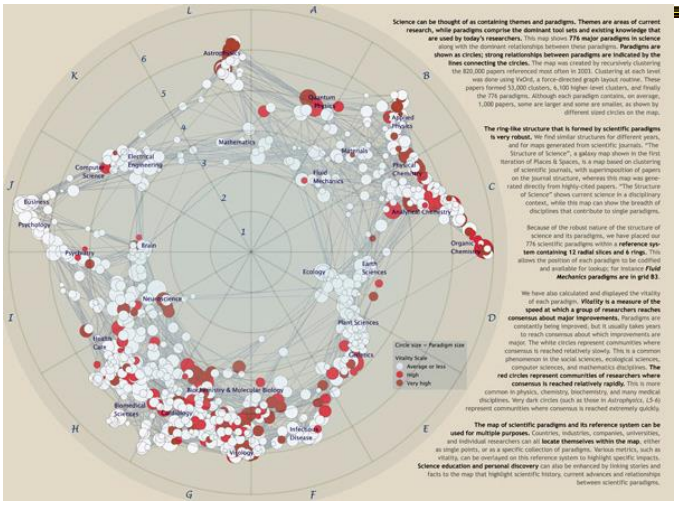
Wonderful Book and Website



K. Börner

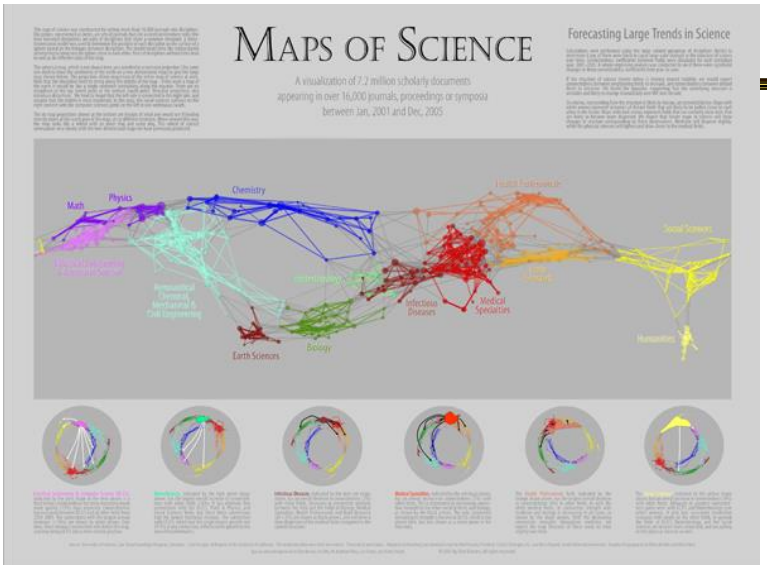
http://scimaps.org

Some Examples



Boyack & Klavans

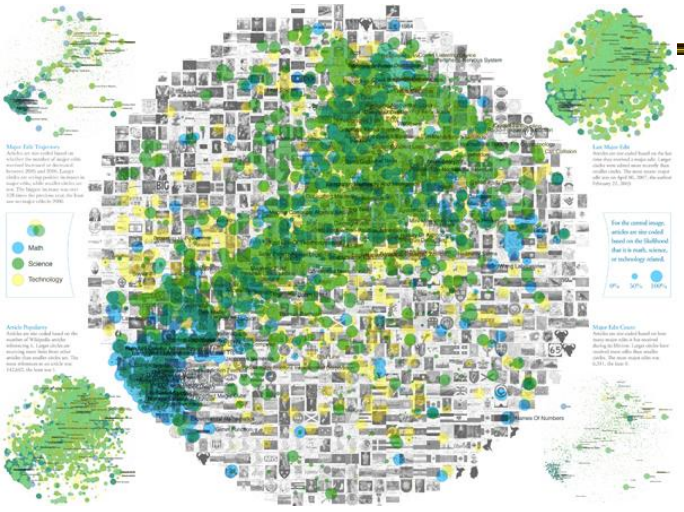
http://scimaps.org/maps/map/map_of_scientific_pa_55/



Klavans & Boyack

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

Science Related Wikipedia Activity



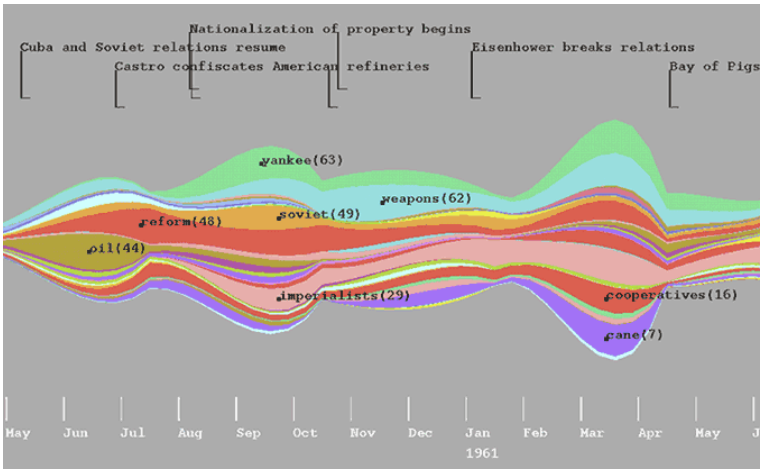
Allgood, Herr, Holloway & Boyack

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

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

Topic Modeling

- Hot topic in text analysis and visualization
- Latent Dirichlet Allocation
- Unsupervised learning
- Produces "topics" evident throughout doc collection, each modeled by sets of words/terms
- Describes how each document contributes to each topic

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

Liu et al
CIKM '09, KDD '10, VAST '10

Representation

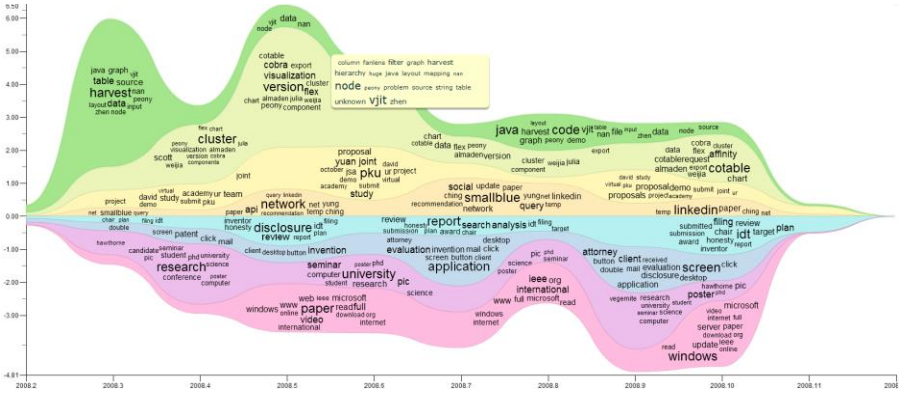
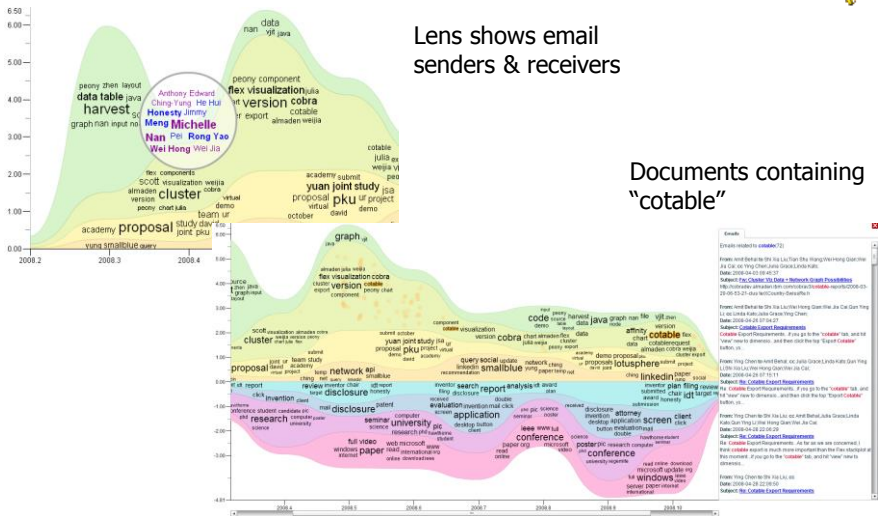
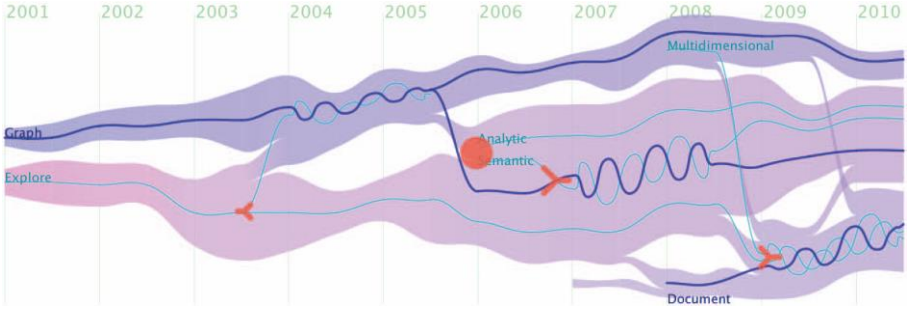


Figure 1. Annotated TIARA-created visual summary of 10,000 emails in the year of 2008. Here, the x-axis encodes the time dimension, the y-axis encodes the importance of each topic. Each layer represents a topic, which is described by a set of keywords. These topic keywords are distributed along the time, summarizing the topic content and the content evolution over time. The tool tip shows the aggregated content of the top-most topic (green one).

Features



TextFlow

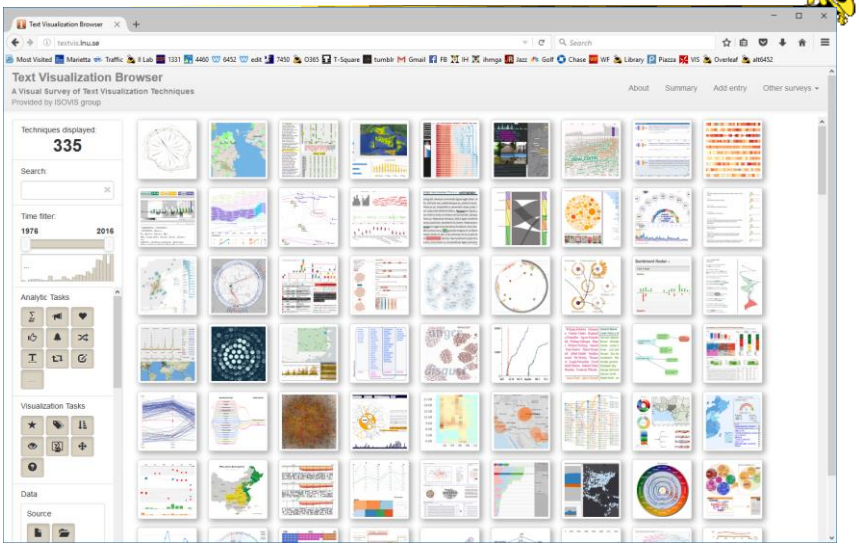


Showing how topics merge and split

Cui et al
TVCG (InfoVis) '11

http://textvis.lnu.se

Text Vis Browser



Learning Objectives

- 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)
- Explain TFIDF
- Describe visual representation used by and contributions of these systems
  - Themail, PaperLens, Jigsaw, ThemeScape/IN-SPIRE, ThemeRiver

Reading

- Meirelles, chapter 6

Upcoming

- Hierarchy and Tree data 1 & 2