

Text and Document Visualization 2



CS 4460 – Intro. to Information Visualization
October 30, 2014
John Stasko

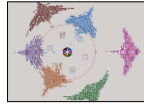
Recall

Example Tasks & Goals

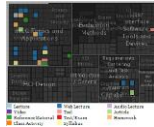


- Which documents contain text on topic XYZ?
- Which documents are of interest to me?
- Are there other documents that are similar to this one (so they are worthwhile)?
- How are different words used in a document or a document collection?
- What are the main themes and ideas in a document or a collection?
- 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.

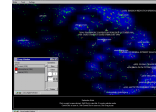
This Week's Agenda



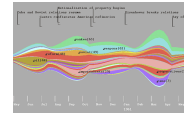
Visualization for IR
Helping search



Visualizing text
Showing words,
phrases, and
sentences



Visualizing document sets
Words & sentences
Analysis metrics
Concepts & themes



Last Time

Fall 2014

CS 4460

3

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

Fall 2014

CS 4460

4

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

Fall 2014

CS 4460

5

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

Fall 2014

CS 4460

6

Vis

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

Do partial overlap in blocks for a smoother appearance

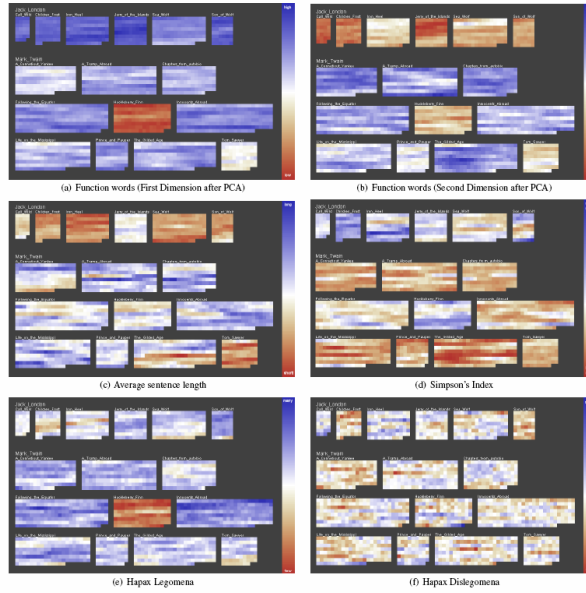


Figure 2: Fingerprints of books of Mark Twain and Jack London. Different measures for authorship attribution are tested. If a measure is able to discriminate between the two authors, the visualizations of the books that are written by the same author will equal each other more than the visualizations 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

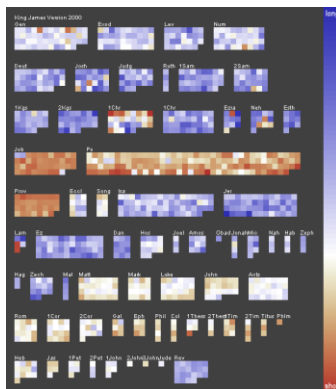


Figure 4: Visual Fingerprint of the Bible. Each pixel represents one chapter of the bible and color is mapped to the average verse length. Interesting characteristics such as the generally shorter verses of the poetry books, the inhomogeneity of the 1. Book of Chronicles or the difference between the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

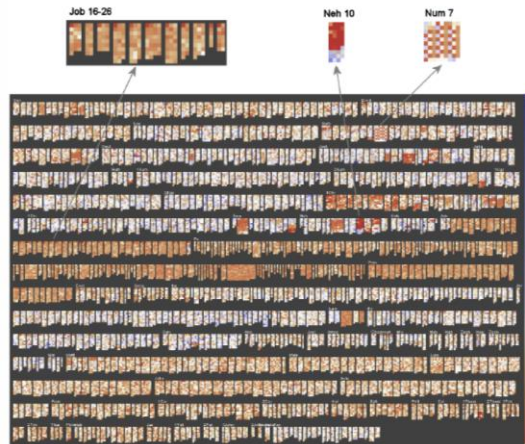


Figure 5: Visual Fingerprint of the Bible. More detailed view on the bible in which each pixel represents a single verse and verses are grouped to chapters. Color is again mapped to verse length. The detailed view reveals some interesting patterns that are camouflaged in the averaged version of fig.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



	Voc. Difficulty	Word Length	Nominal Forms	Sent. Length	Compl. Sent. Struc.
(a) The intention of TileBars [9] is to provide a compact but yet meaningful representation of Information Retrieval results, whereas the FeatureLens technique, presented in [5], was designed to explore interesting text patterns which are suggested by the system, find meaningful co-occurrences of them, and identify their temporal evolution.					
(b) This includes aspects like ensuring contextual coherency, 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)

Interface

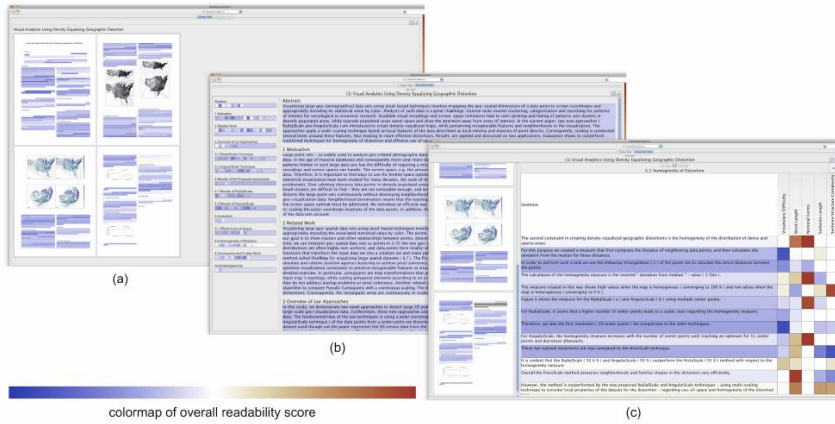


Figure 3: Screenshot of the VisRA tool on 3 different aggregation levels. (a) Corpus View (b) Block View (c) Detail View. To display single features, the colormap is generated as described in section 3.4 and figure 2.

Their Paper (Before & After)

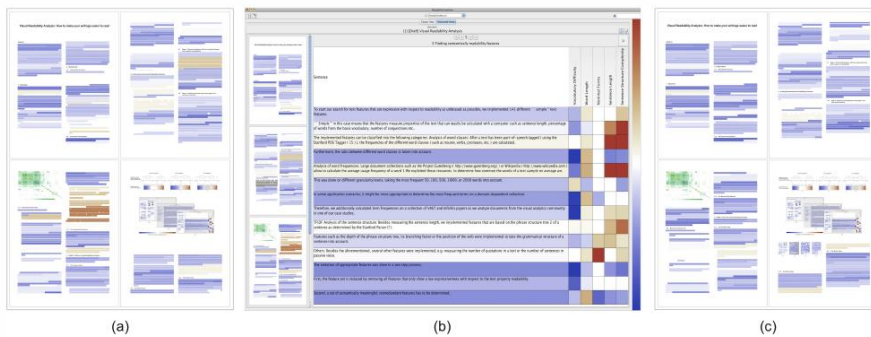


Figure 6: Revision of our own paper. (a) The first four pages of the paper as structure thumbnails before the revision. (b) Detail view for one of the sections. (c) Structure thumbnails of the same pages after the revision.

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

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

Strobelt et al
TVCG (InfoVis) '09

Fall 2014

CS 4460

15

Representation



Layout algorithm searches for empty space rectangles to put things

Fall 2014

CS 4460

16

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

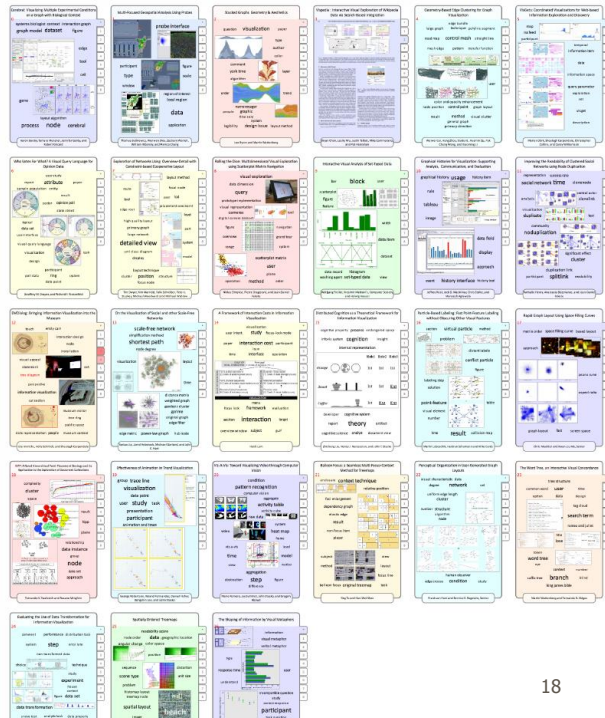
Fall 2014

CS 4460

17

InfoVis '08
Proceedings

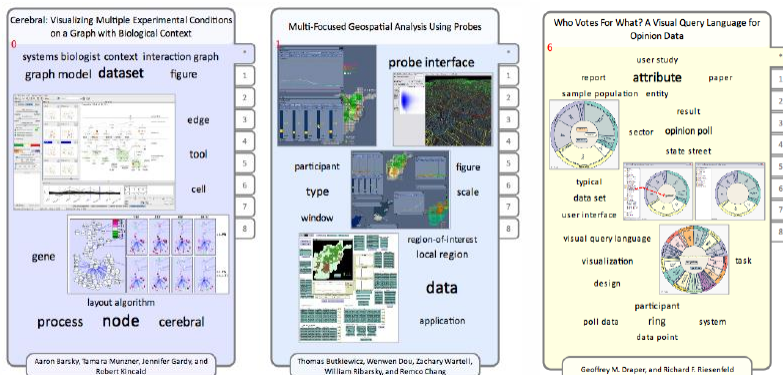
Fall 2014



18



Zooming In



Fall 2014

CS 4460

19

Bohemian Bookshelf



Serendipitous browsing



Thudt et al
CHI '12

Fall 2014

CS 4460

20

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?

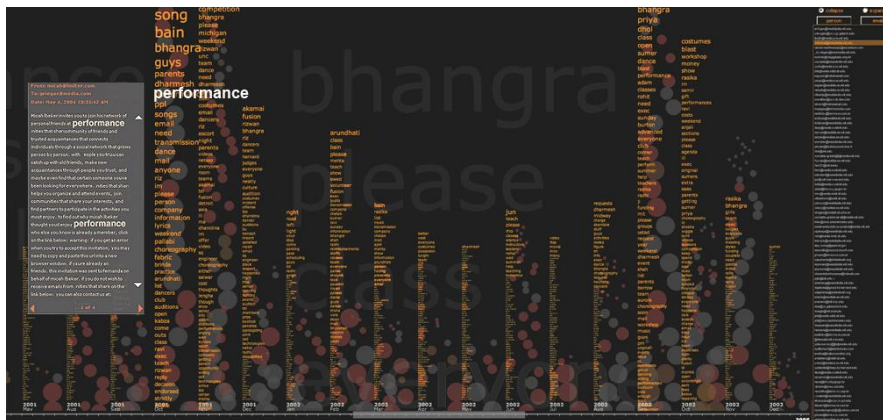
Fall 2014

CS 4460

Viégas, Golder & Donath
CHI '06

21

Interface



Fall 2014

CS 4460

22

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

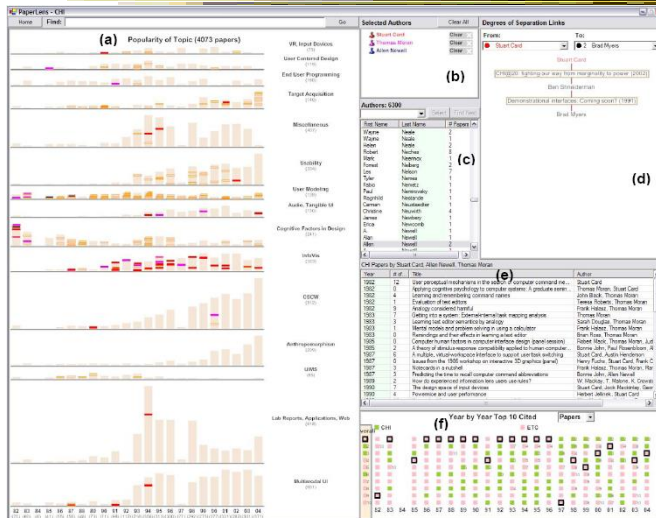
Fall 2014

CS 4460

25

PaperLens

Video



- Popularity of topic
- Selected authors
- Author list
- Degrees of separation of links
- Paper list
- Year-by-year top ten cited papers/ authors – can be sorted by topic

Fall 2014

CS 4460

26



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 panel shows the distribution of papers (in logarithmic scale) over time, grouped by topics. Users can see which topics have their 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.

Fall 2014

CS 4460

27

More Document Info



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

Fall 2014

CS 4460

28

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

Fall 2014

CS 4460

29

Document View



The screenshot shows a software interface titled "Document View". At the top, there is a word cloud with terms like "analysis", "information", "interaction", "visualization", and "systems". Below the word cloud is a "Documents" list on the left, with one document selected. The main area displays the text of the selected document, with several words highlighted in yellow and blue, indicating identified entities. Labels with arrows point to these features: "Wordcloud overview" points to the top word cloud; "Document summary" points to the text area; "Doc List" points to the list on the left; and "Selected document's text with entities identified" points to the highlighted text in the main area.

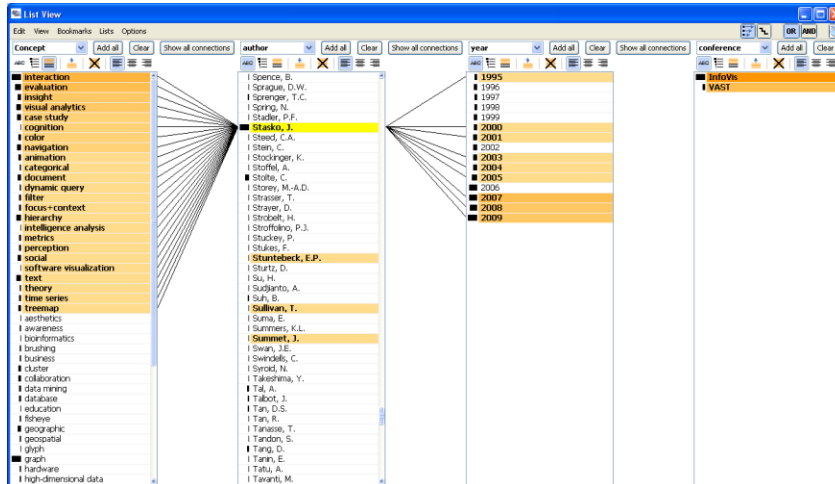
Fall 2014

CS 4460

30

List View

Entities listed by type



Fall 2014

CS 4460

31

Document Cluster View

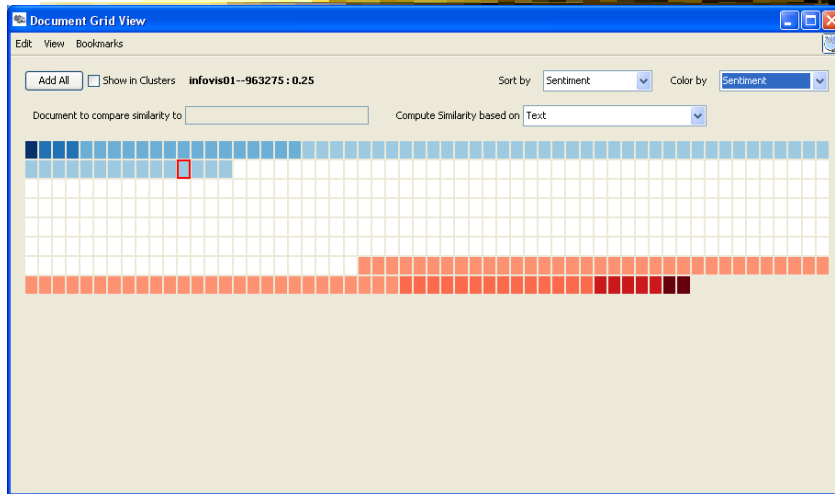


Fall 2014

CS 4460

32

Document Grid View



Here showing sentiment analysis of docs

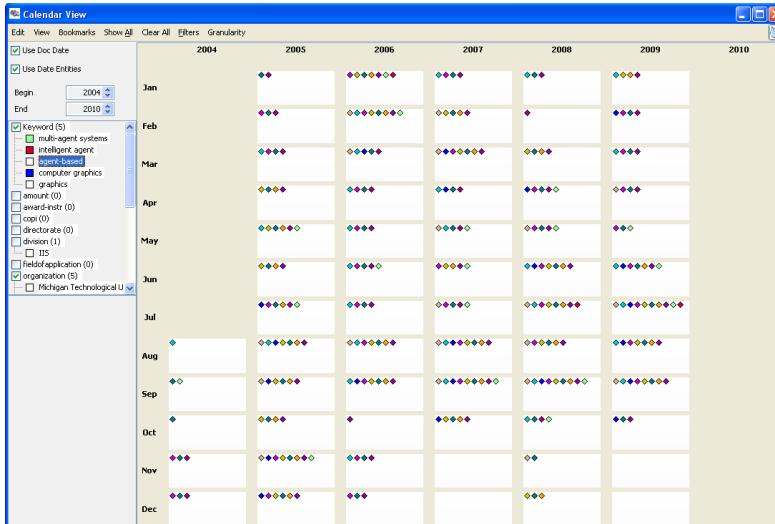
Fall 2014

CS 4460

33

Calendar View

Temporal context of entities & docs



Fall 2014

CS 4460

Video

34

Jigsaw



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

Fall 2014

CS 4460

35

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

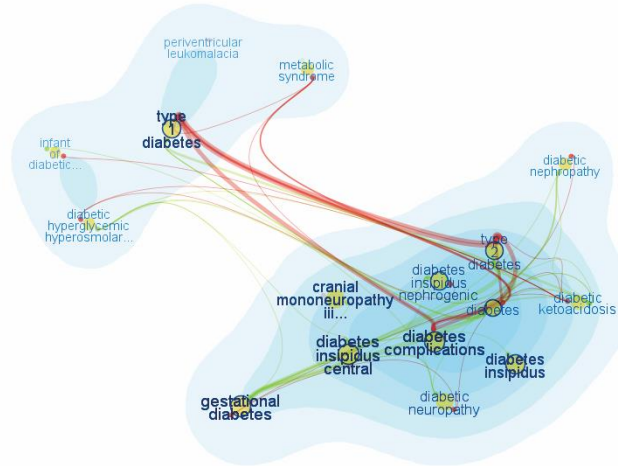
Cao et al
TVCG (InfoVis) '10

Fall 2014

CS 4460

36

Visualization



Fall 2014

CS 4460

37

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

Fall 2014

CS 4460

38

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

Fall 2014

CS 4460

39

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

Fall 2014

CS 4460

40

Wiggle



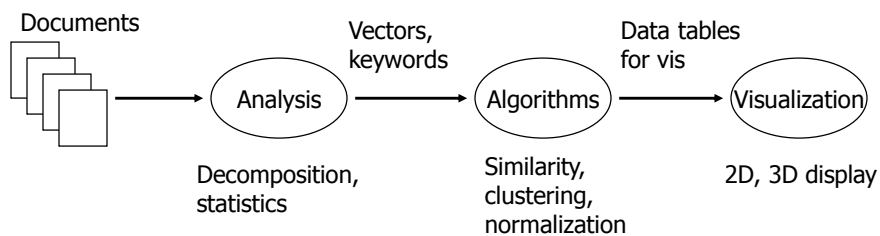
- Not all terms or words are equally useful
- Often apply TFIDF
 - Term frequency, inverse document frequency
- Weight of a word goes up if it appears often in a document, but not often in the collection

Fall 2014

CS 4460

41

Process



Fall 2014

CS 4460

42

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

Fall 2014

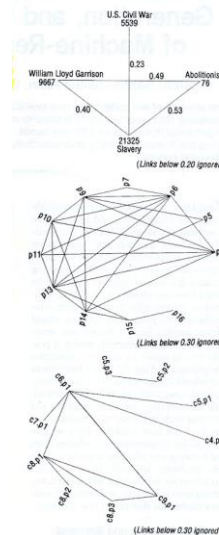
CS 4460

43

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

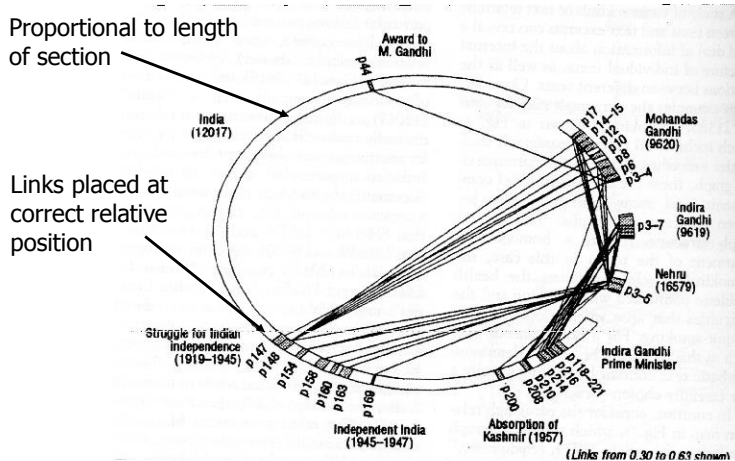


Fall 2014

CS 4460

44

Improved Design



Fall 2014

CS 4460

45

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

Fall 2014

CS 4460

46

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 shown
- Also visualize how documents are similar or different

Fall 2014

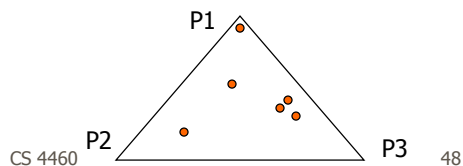
CS 4460

47

Technique



- Represent points of interest as vertices on a 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

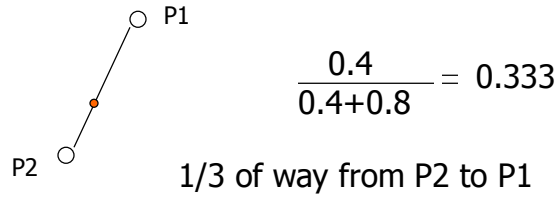


Fall 2014

Algorithm



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



Fall 2014

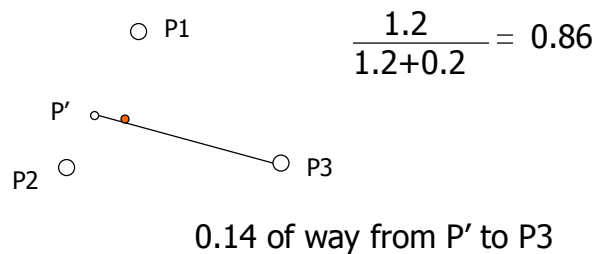
CS 4460

49

Algorithm



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

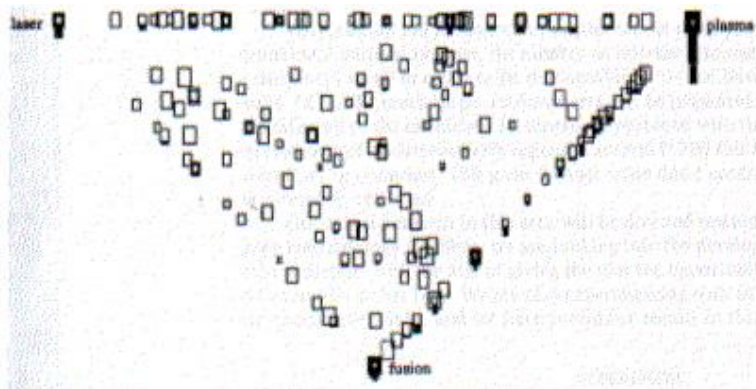


Fall 2014

CS 4460

50

Sample Visualization



Fall 2014

CS 4460

51

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

Fall 2014

CS 4460

52

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

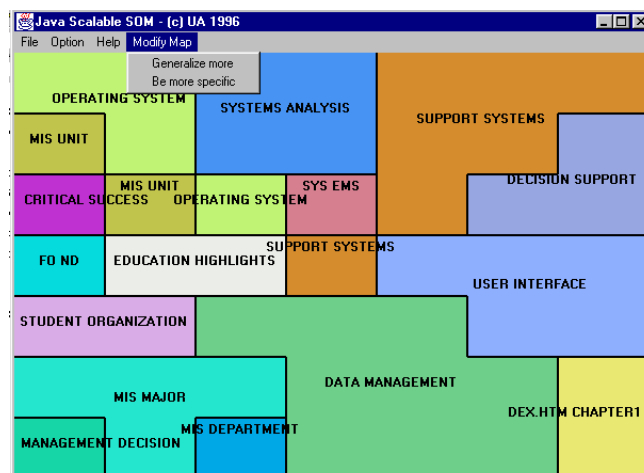
Lin
Visualization '92

Fall 2014

CS 4460

55

Map Display of SOM



Fall 2014

CS 4460

56

Map Attributes



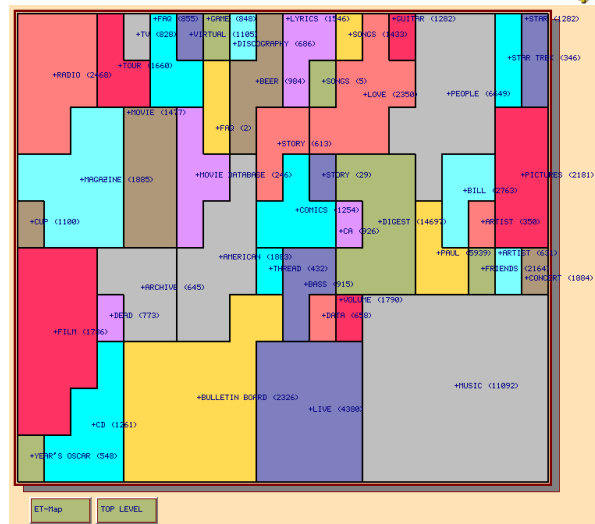
- Different, colored areas correspond to different concepts in collection
- Size of area corresponds to its relative importance in set
- Neighboring regions indicate commonalities in concepts
- Dots in regions can represent documents

Fall 2014

CS 4460

57

More Maps



Fall 2014

CS 4460

58

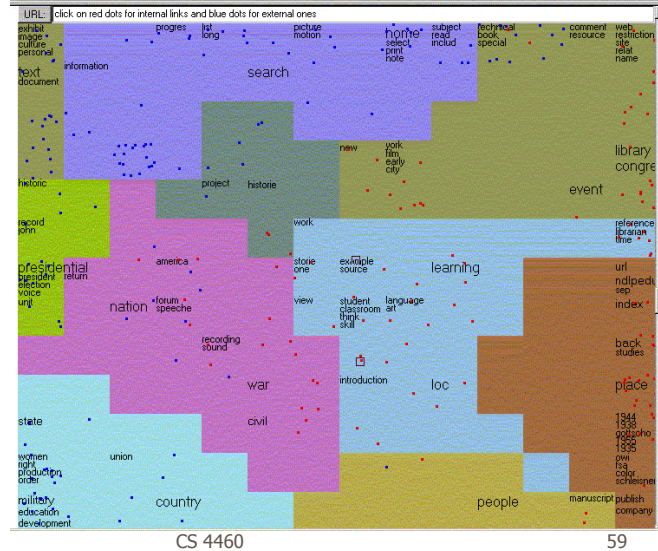
More Maps



Interactive demos

Xia Lin

Fall 2014



Work at PNNL

<http://www.pnl.gov/infviz>



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

Wise et al
InfoVis '95

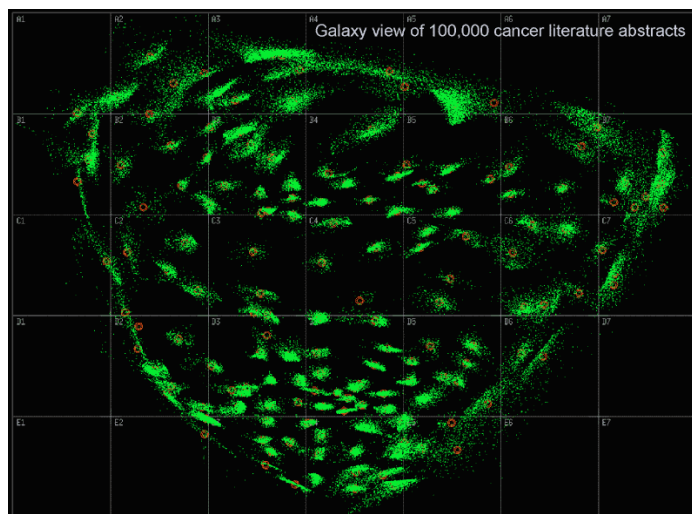
Fall 2014

CS 4460

60

Galaxies

Presentation of documents where similar ones cluster together



Fall 2014

CS 4460

61

Themespaces



- 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

Fall 2014

CS 4460

62

Related Topic



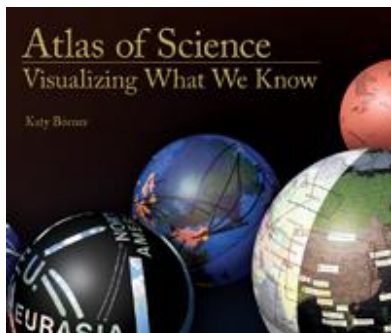
- Maps of Science
- Visualize the relationships of areas of science, emerging research disciplines, the impact of particular researchers or institutions, etc.
- Often use documents as the “input data”

Fall 2014

CS 4460

65

Wonderful Book and Website



K. Börner



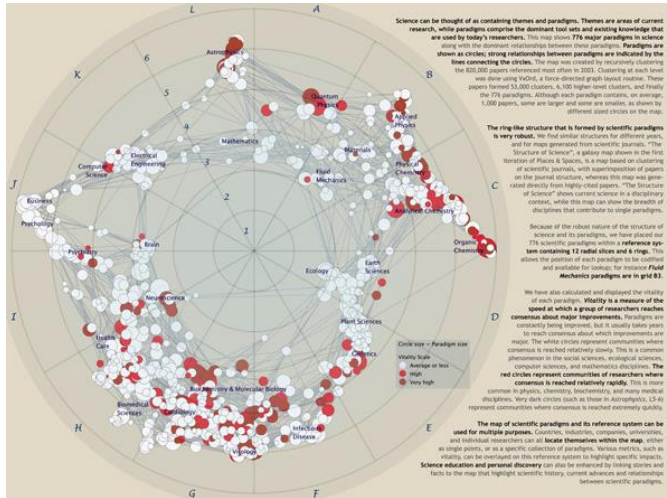
<http://scimaps.org>

Fall 2014

CS 4460

66

Some Examples



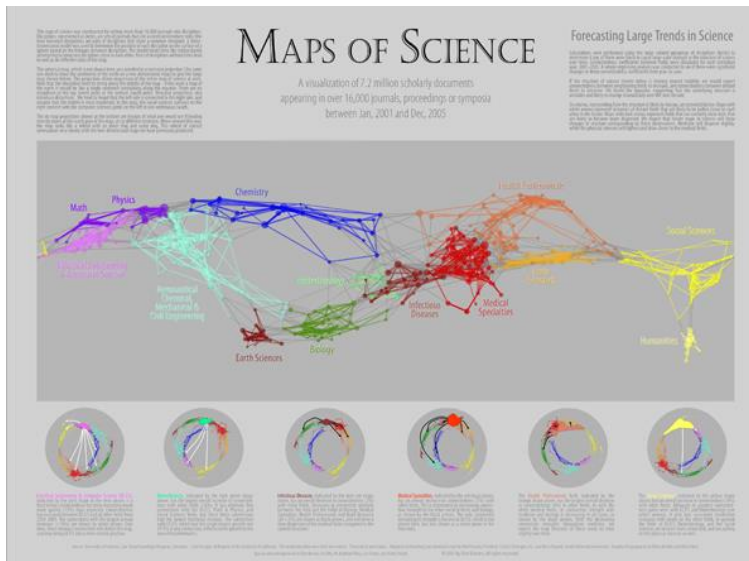
Boyack & Klavans

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

Fall 2014

CS 4460

67



Klavans & Boyack

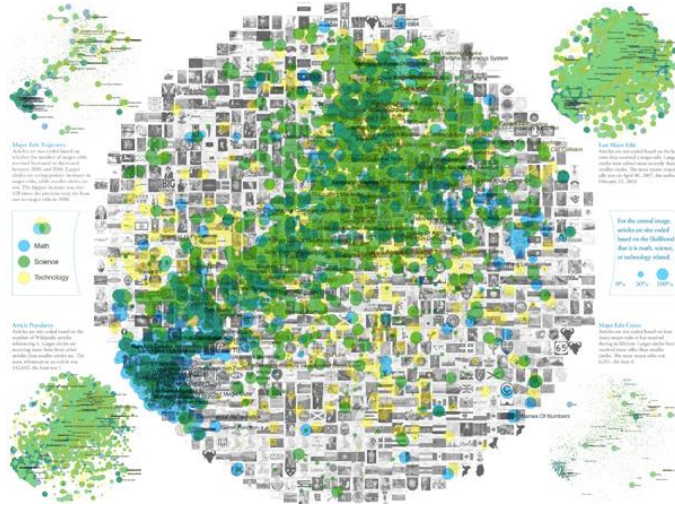
http://scimaps.org/maps/map/maps_of_science_fore_50/

Fall 2014

CS 4460

68

Science Related Wikipedia Activity



Allgood,
Herr,
Holloway &
Boyack

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

Fall 2014

CS 4460

69

Temporal Issues

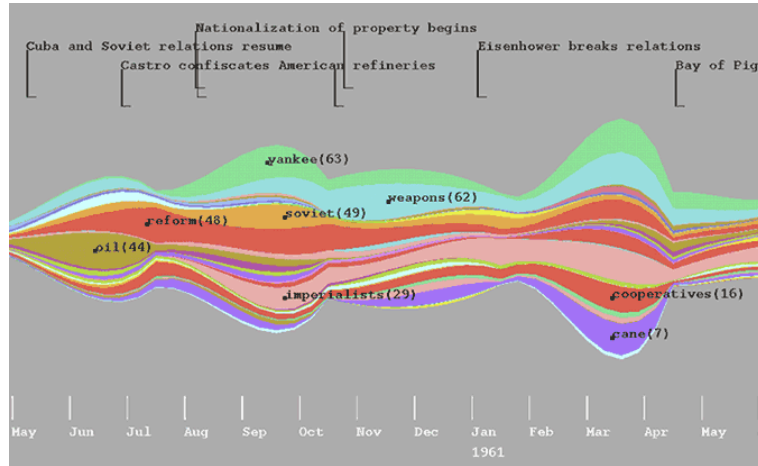
- Semantic map gives no indication of the chronology of documents
- Can we show themes and how they rise or fall over time?

Fall 2014

CS 4460

70

ThemeRiver



Fall 2014

CS 4460

Havre, Hetzler, & Nowell
InfoVis '00

71

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

Fall 2014

CS 4460

72

More Information



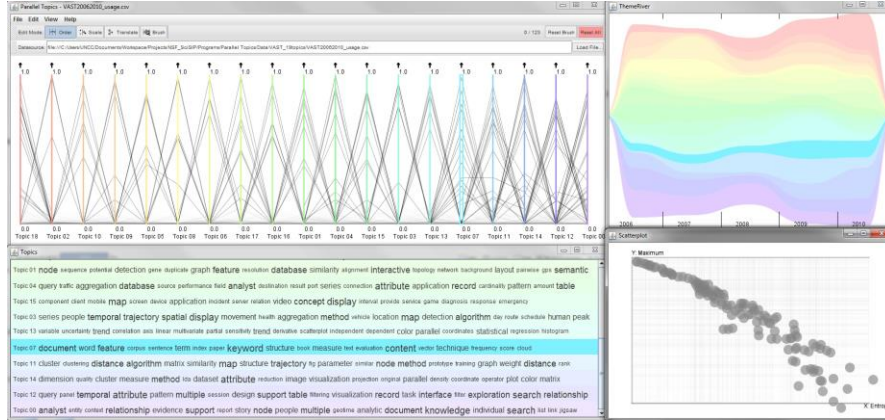
- What's in the bands?
- Analysts may want to know about what each band is about

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

ParallelTopics



Dou et al
VAST '11

Fall 2014

CS 4460

79

Upcoming



- Interaction
 - Reading
 - Now You See It*, chapter 4
 - Munzner chapters 11 and 13
- Commercial Systems
 - Reading

Fall 2014

CS 4460

80