Text and Document Visualization
Part 2

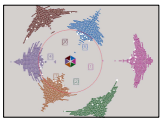
CS 4460 - Information Visualization

Spring, 2019

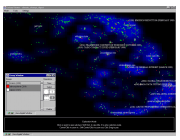
Alex Endert

This Week's Agenda

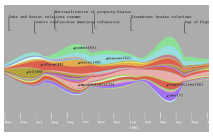
Visualization for IR
Helping search



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

    Need analytics to help create the vis -> Visual Analytics

Various Document Metrics

- Goal is to visualize metrics about entire documents for comparison

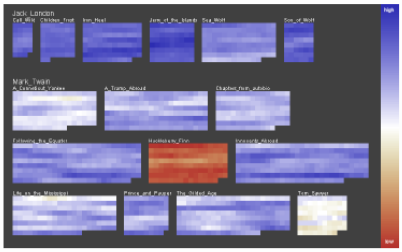
- Different variables for literary analysis
  - Average word length
  - Syllables per word
  - Average sentence length
  - Percentage of nouns, verbs, adjectives
  - Frequencies of specific words
  - Hapax Legomena – number of words that occur once

Keim & Oelke
VAST '07

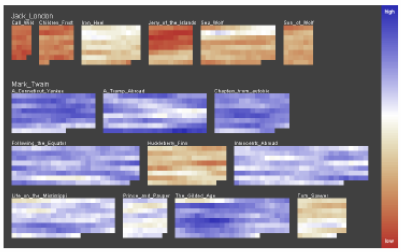
Vis

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

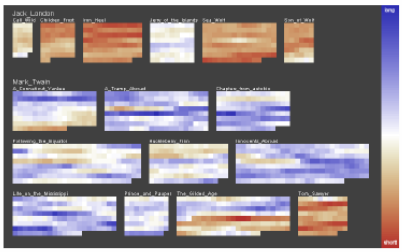
Do partial overlap in blocks for a smoother appearance



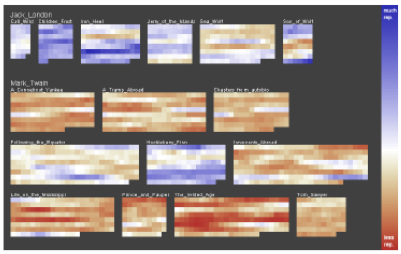
(a) Function words (First Dimension after PCA)



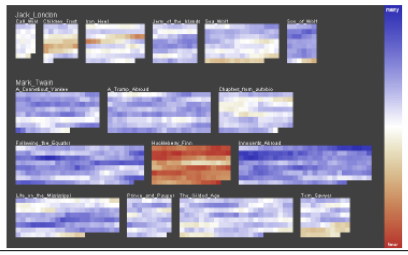
(b) Function words (Second Dimension after PCA)

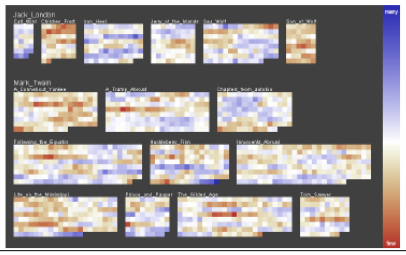


(c) Average sentence length



(d) Simpson's Index

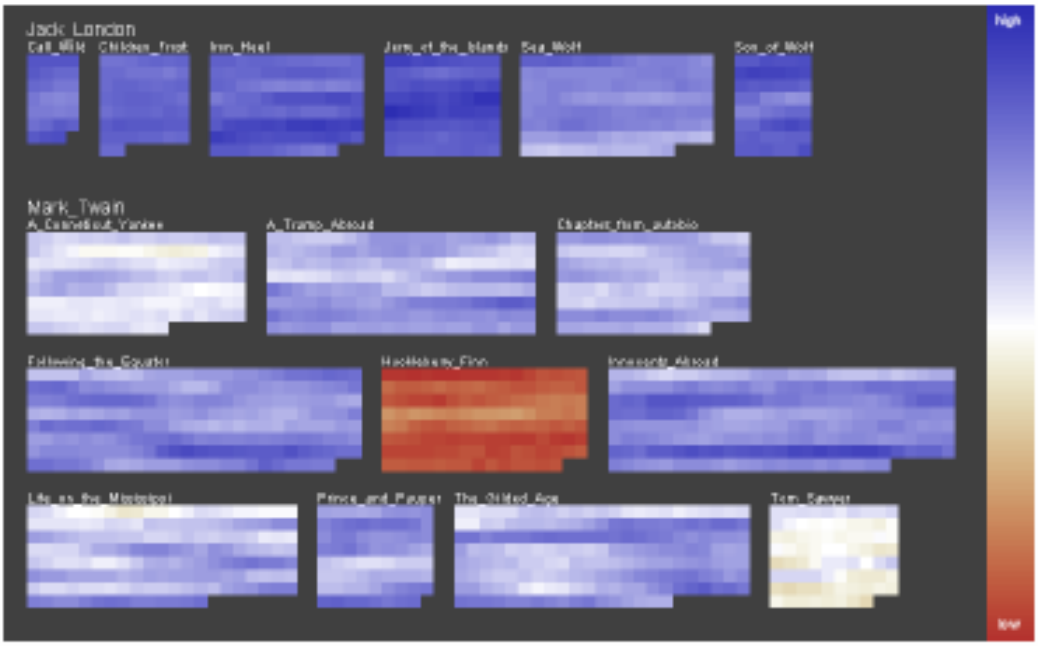




a closer look

anyone a Mark Twain fan?

Was Huckleberry Finn written by someone else?



(a) Function words (First Dimension after PCA)

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

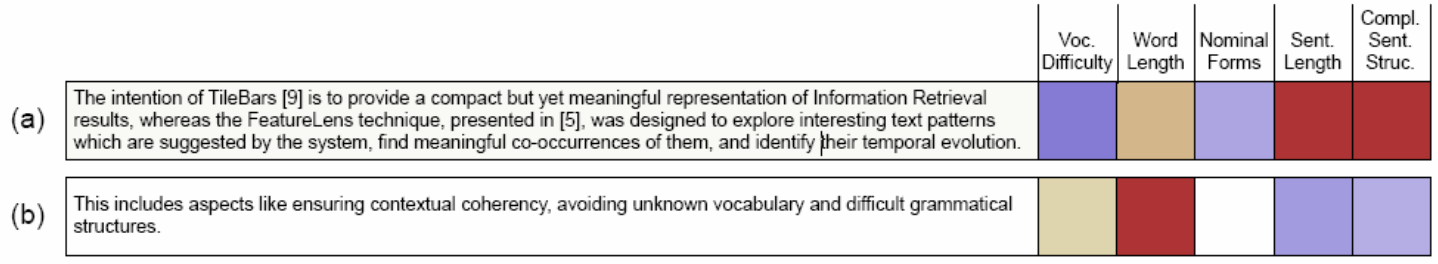
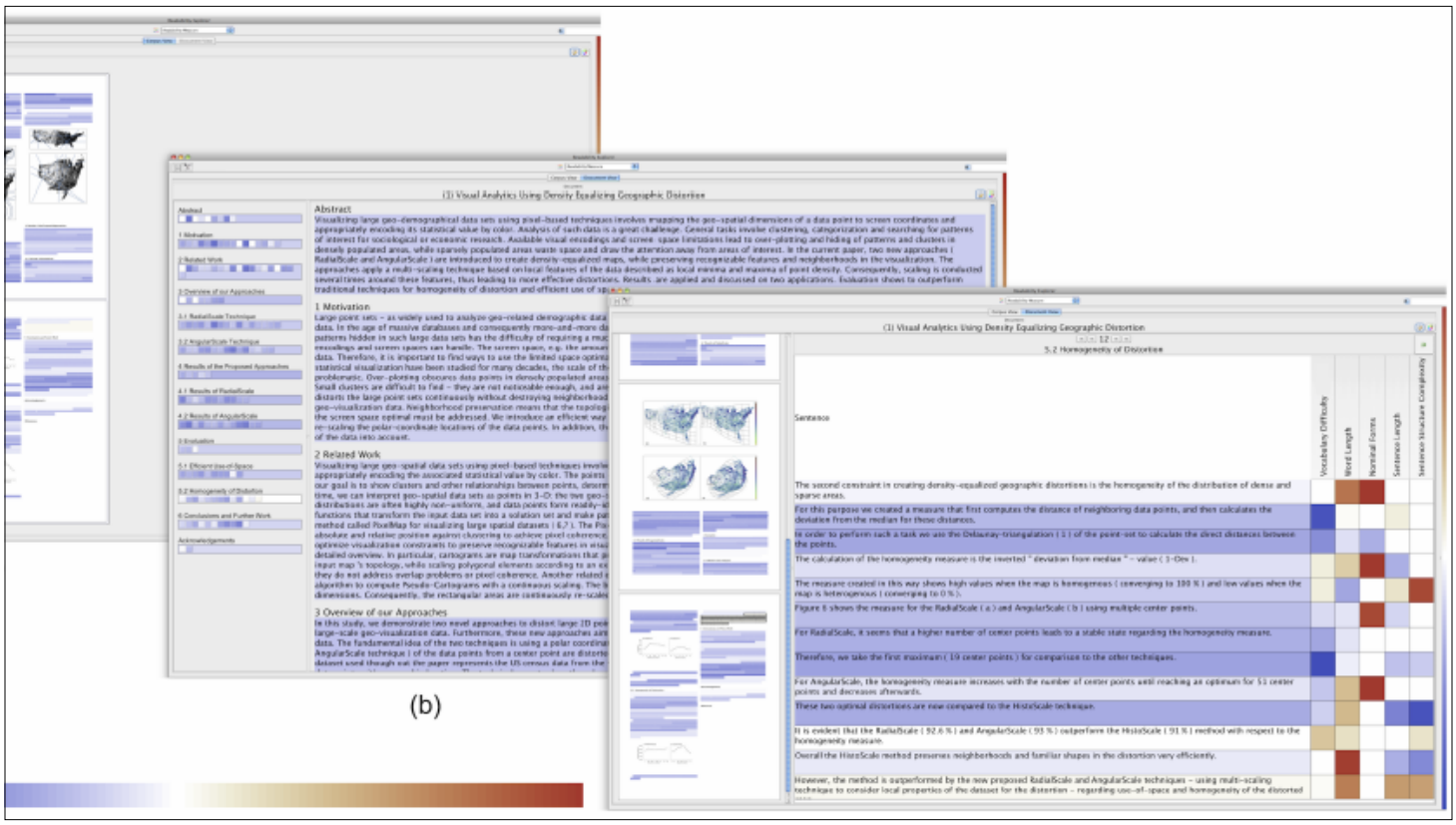


Figure 5: Two example sentences whose overall readability score is about the same. The detail view reveals the different reasons why the sentences are difficult to read.

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



Their Own Paper (Before & After)

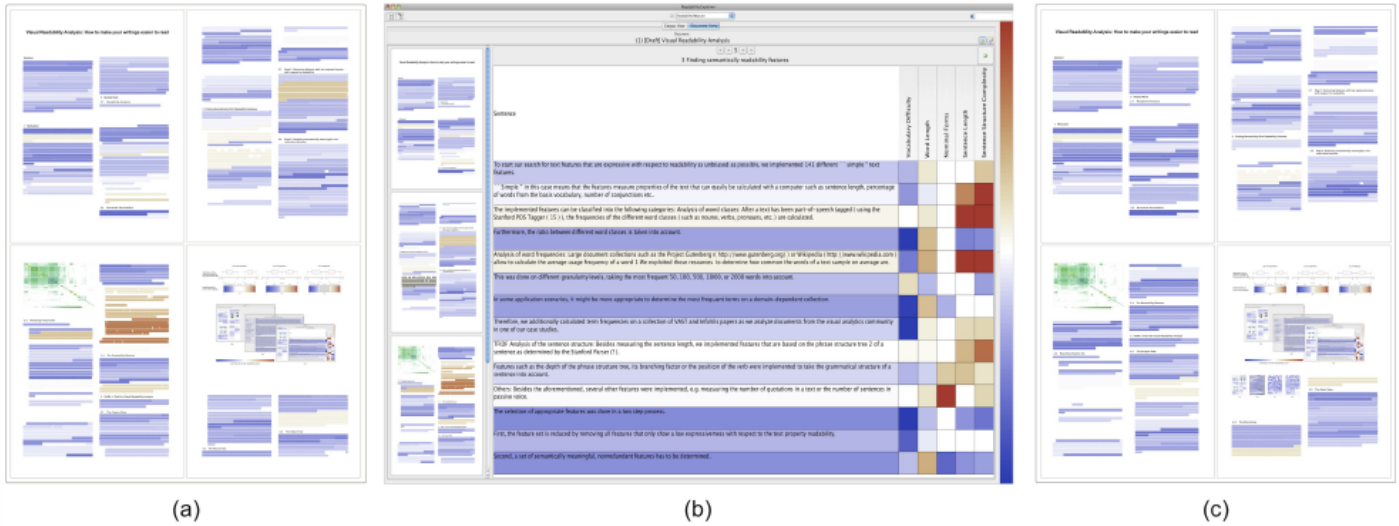


Figure 6: Revision of our own paper. (a) The first four pages of the paper as structure thumbnails before the revision. (b) Detail view for one of the sections. (c) Structure thumbnails of the same pages after the revision.

Document Cards

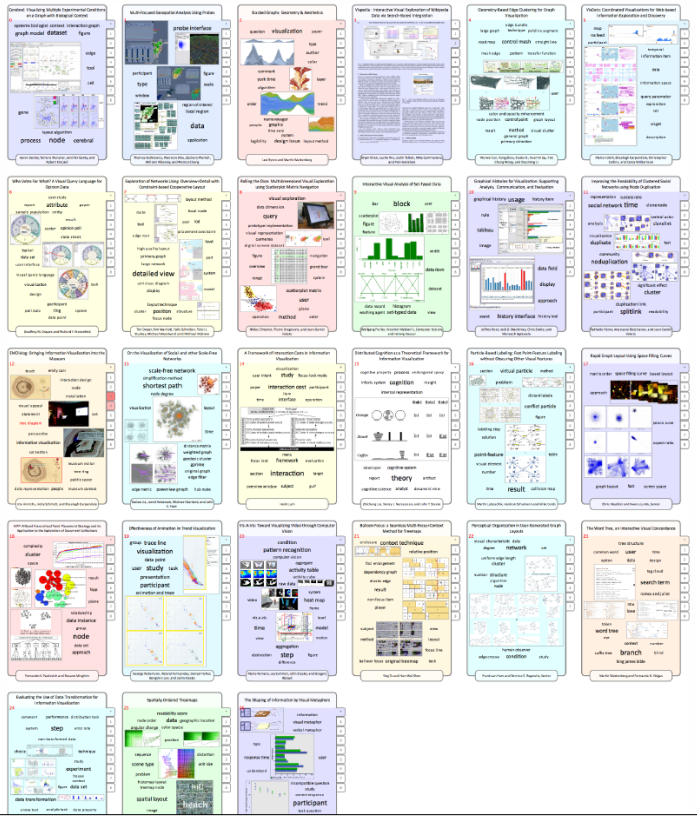
- Compact visual representation of a document
- Show key terms and important images

Strobelt et al
TVCG (InfoVis) ‘09

Representation

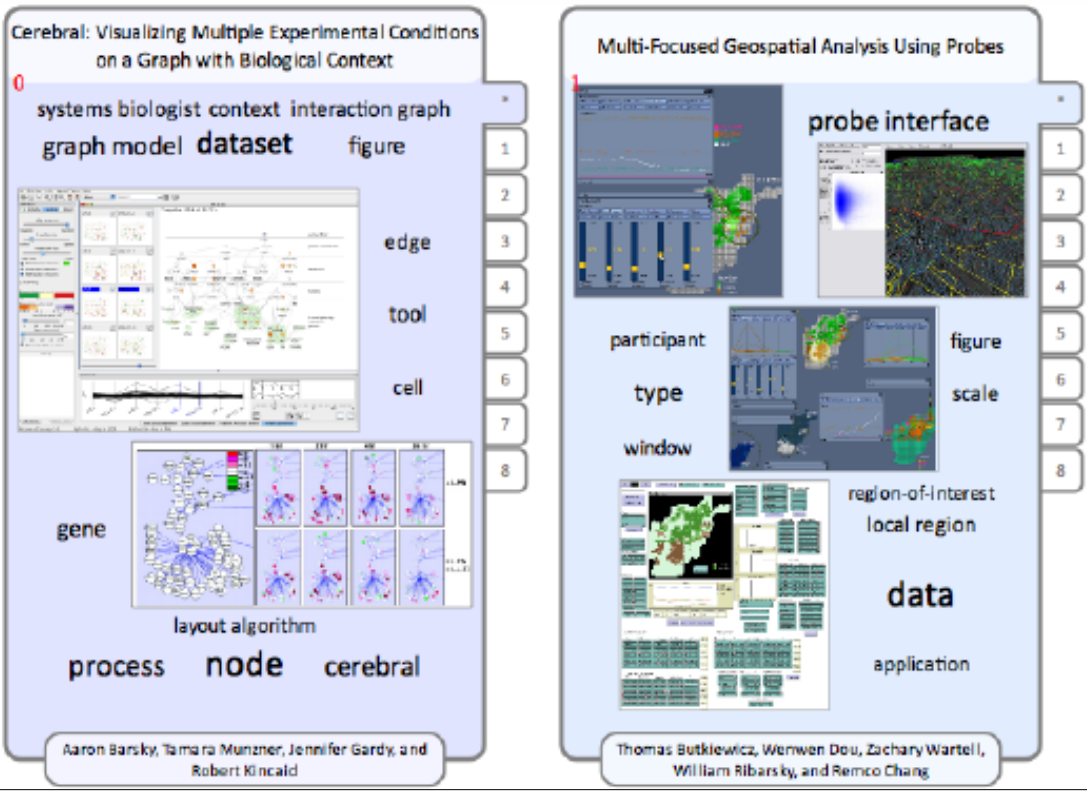
Layout algorithm searches for empty space rectangles to put things

InfoVis '08 Proceedings



endert@gatech.edu

Zooming In



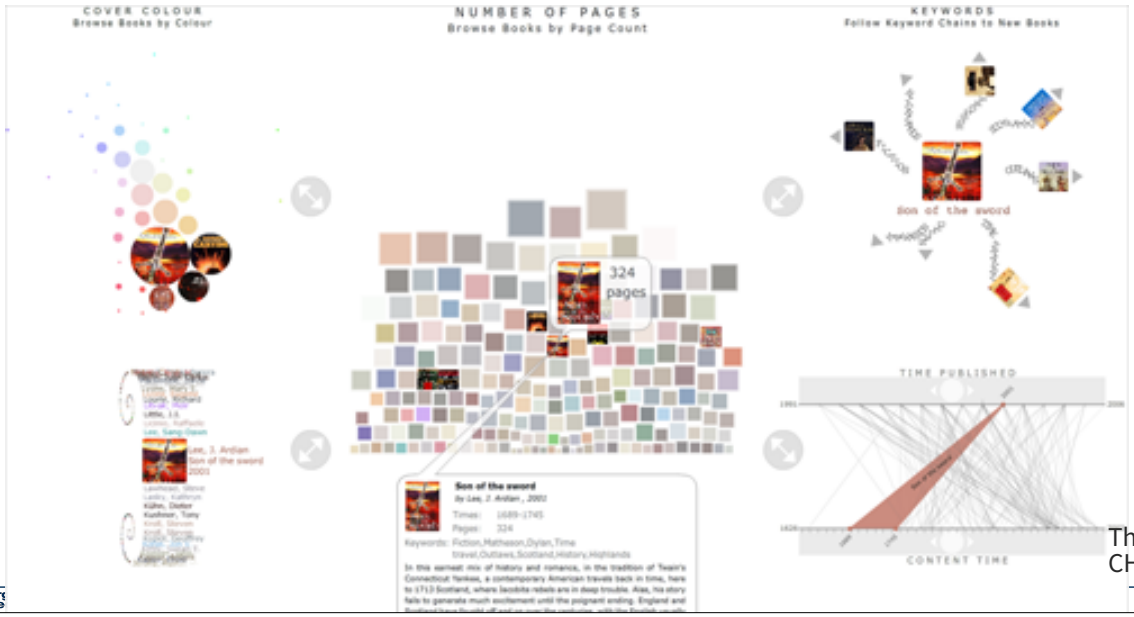
endert@gatech.edu

Bohemian Bookshelf

really about casual browsing

Video

Serendipitous browsing



Thudt et al
CHI '12

Jigsaw

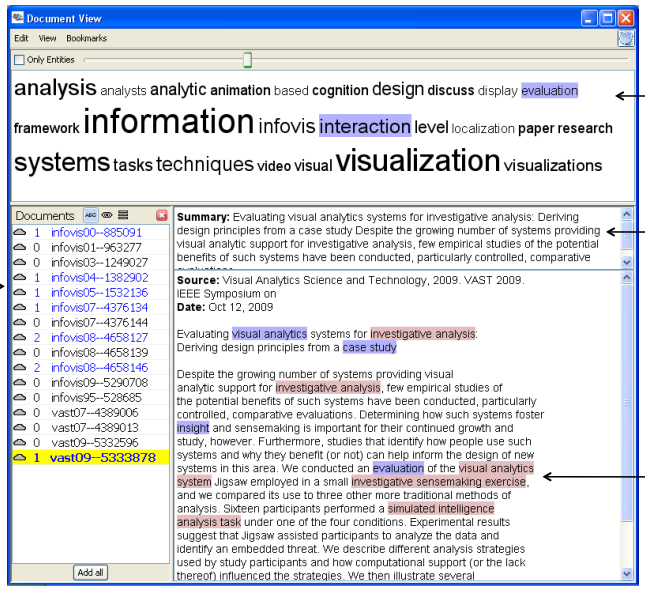
- Targeting sense-making scenarios

In contrast to Bohemian bookshelf

- 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 View



Wordcloud overview

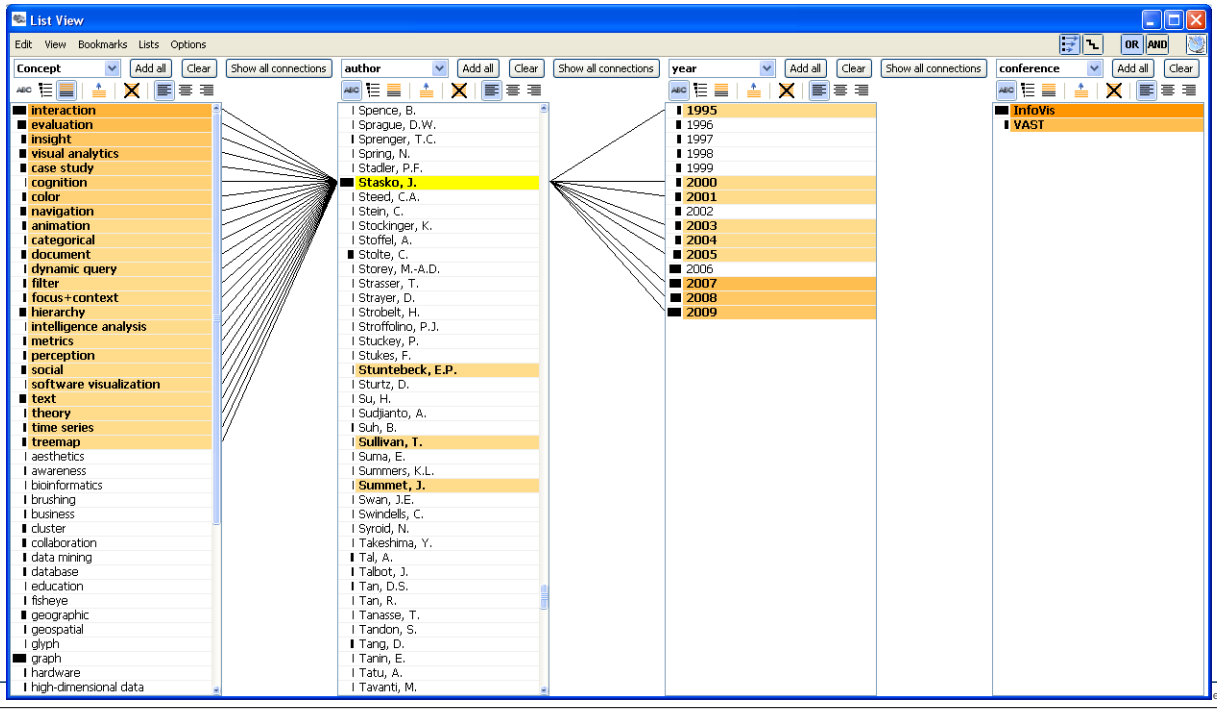
Document summary

Doc List

Selected document's text with entities identified

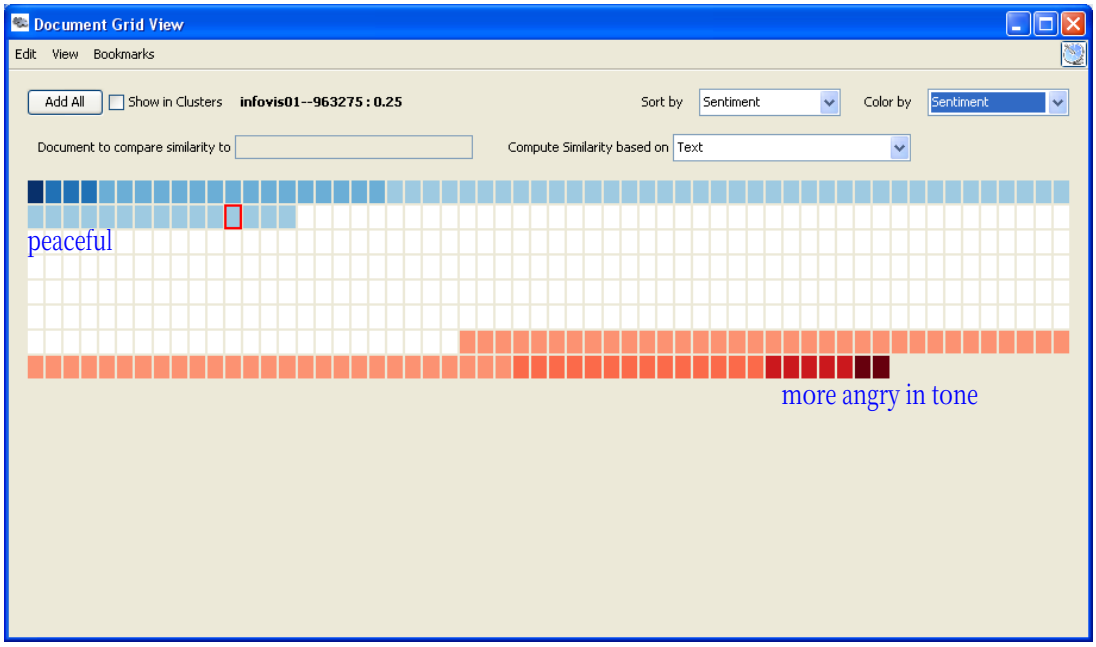
List View

Entities listed by type



Document Cluster View

Document Grid View



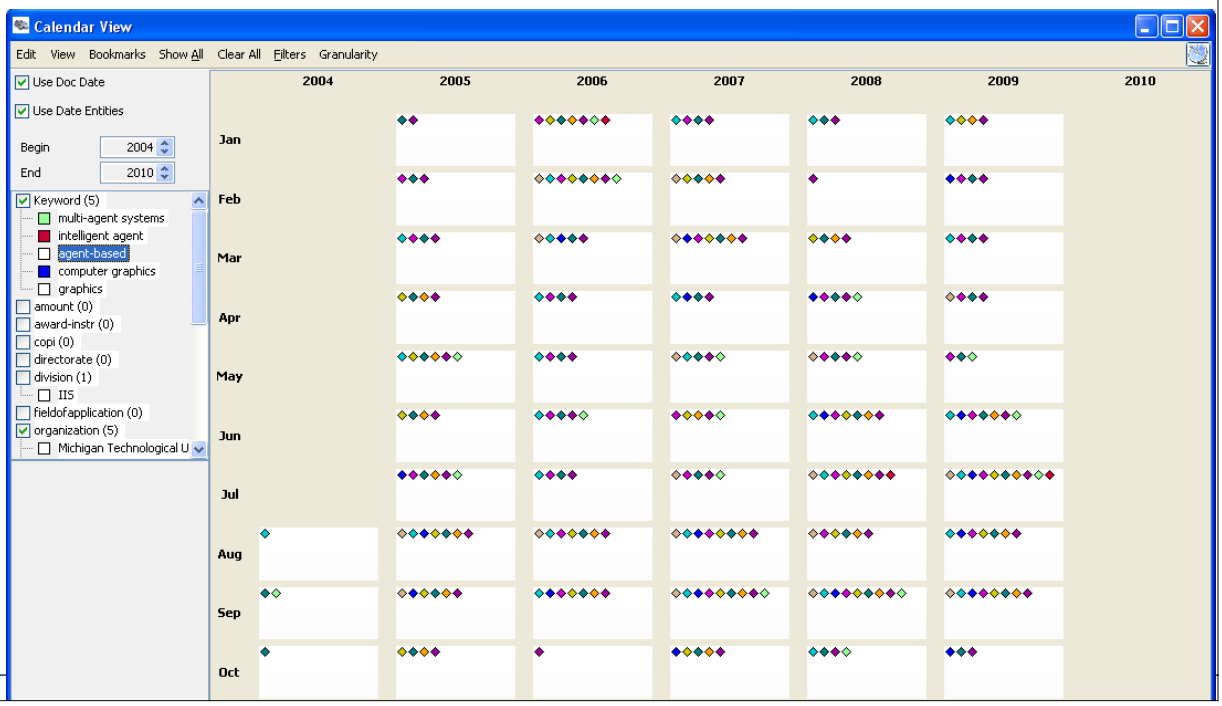
Here showing sentiment analysis of docs

Calendar View

Temporal context of entities & docs

represent if a document has a time step.

a nice temporal view of what different "entities" are doing at different times.



Jigsaw

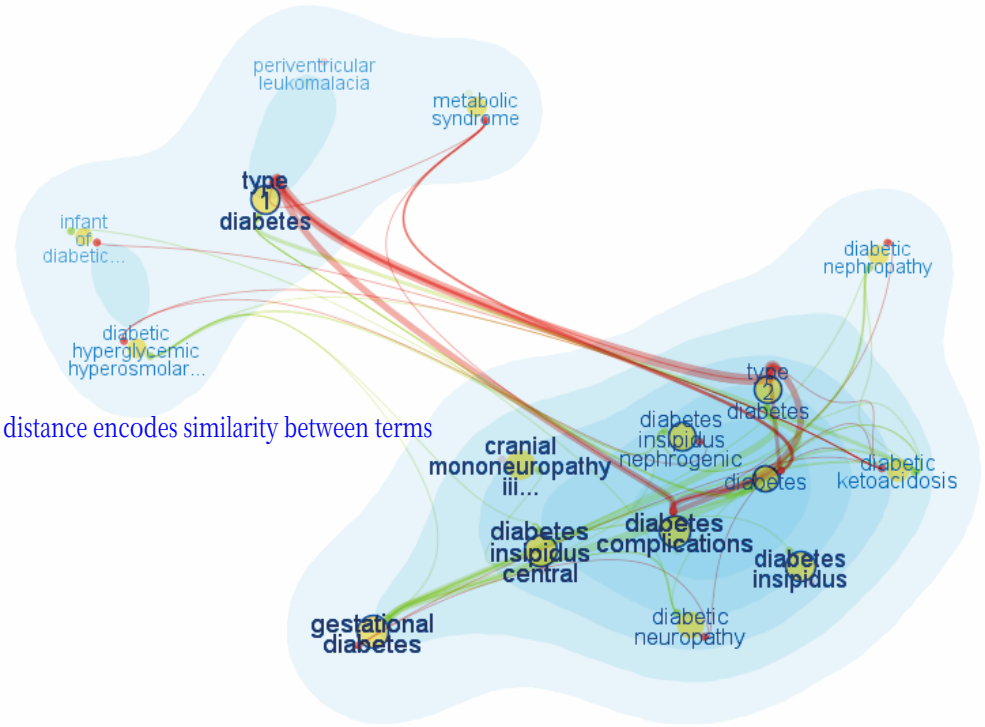
- Much more to come on Visual Analytics week…
- For now, let's try one of the views:
  - ListView: http://www.iilabgt.org/listview/

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

FacetAtlas



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) cosine-similarity
  - 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

Not all words are equal

- 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
- words that show up frequently over the entire document collection may not be that important

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

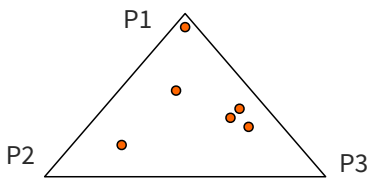
Focus

- Points of Interest
  - Terms or keywords that are of interest to user

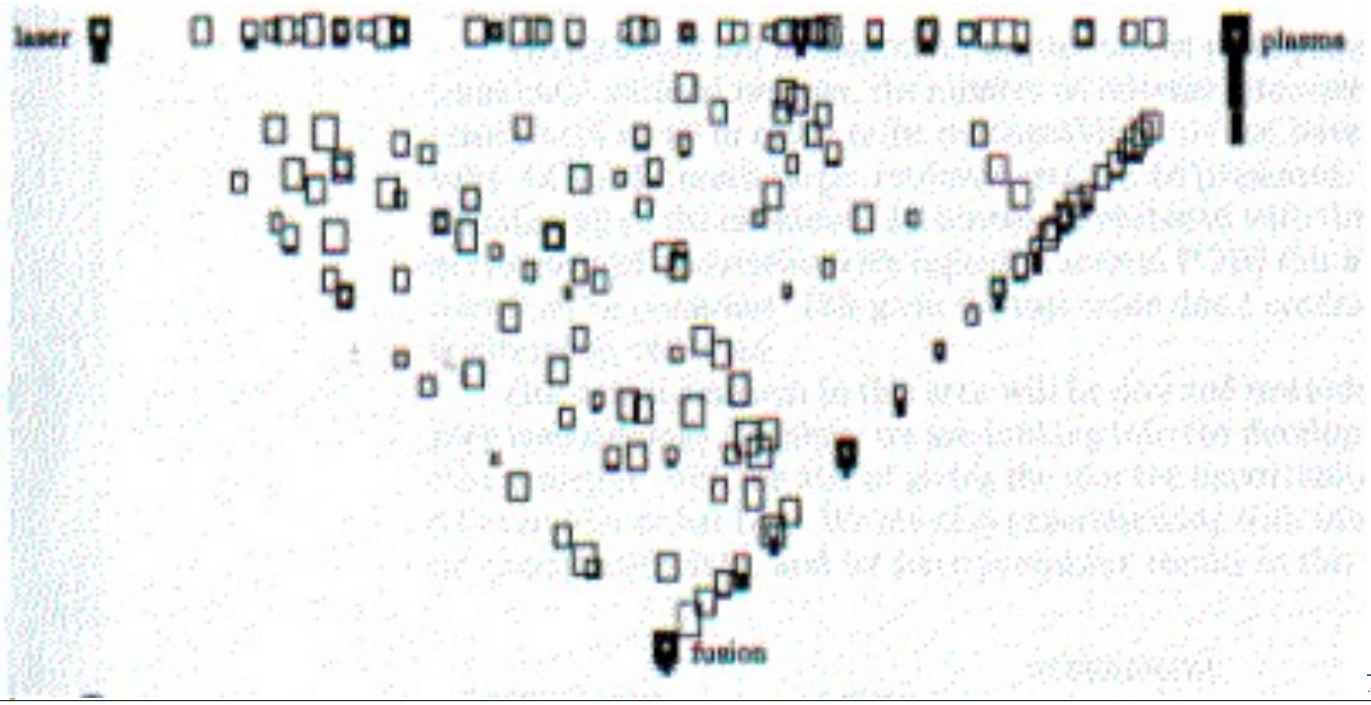
    Example: cooking, pies, apples
- Want to visualize a document collection where each document's relation to points of interest is show
- Also visualize how documents are similar or different

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



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

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

Work at PNNL

http://www.pnl.gov/infoviz

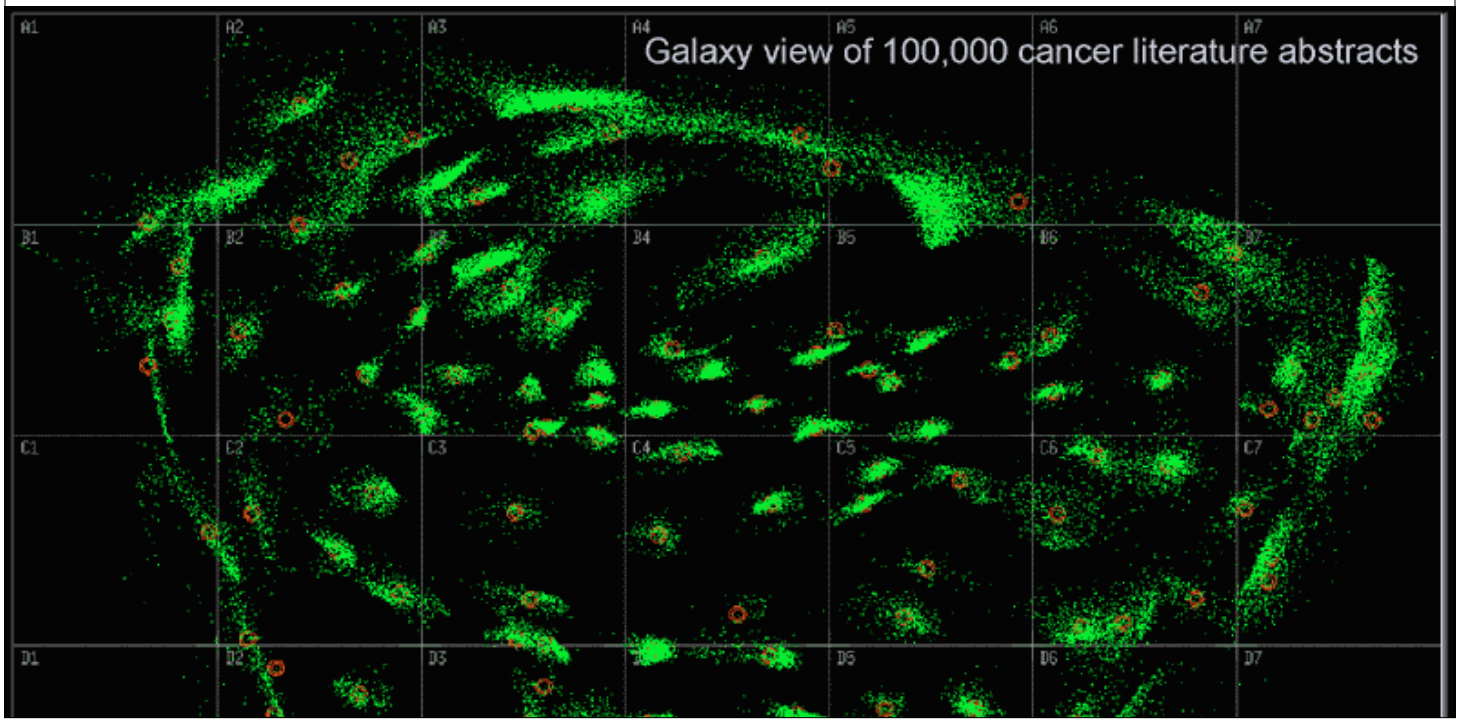
- Group has developed a number of visualization techniques for document collections
  - Galaxies
  - Themescapes
  - ThemeRiver
  - …

Wise et al
InfoVis '95

Galaxies

Presentation of documents where similar ones cluster together

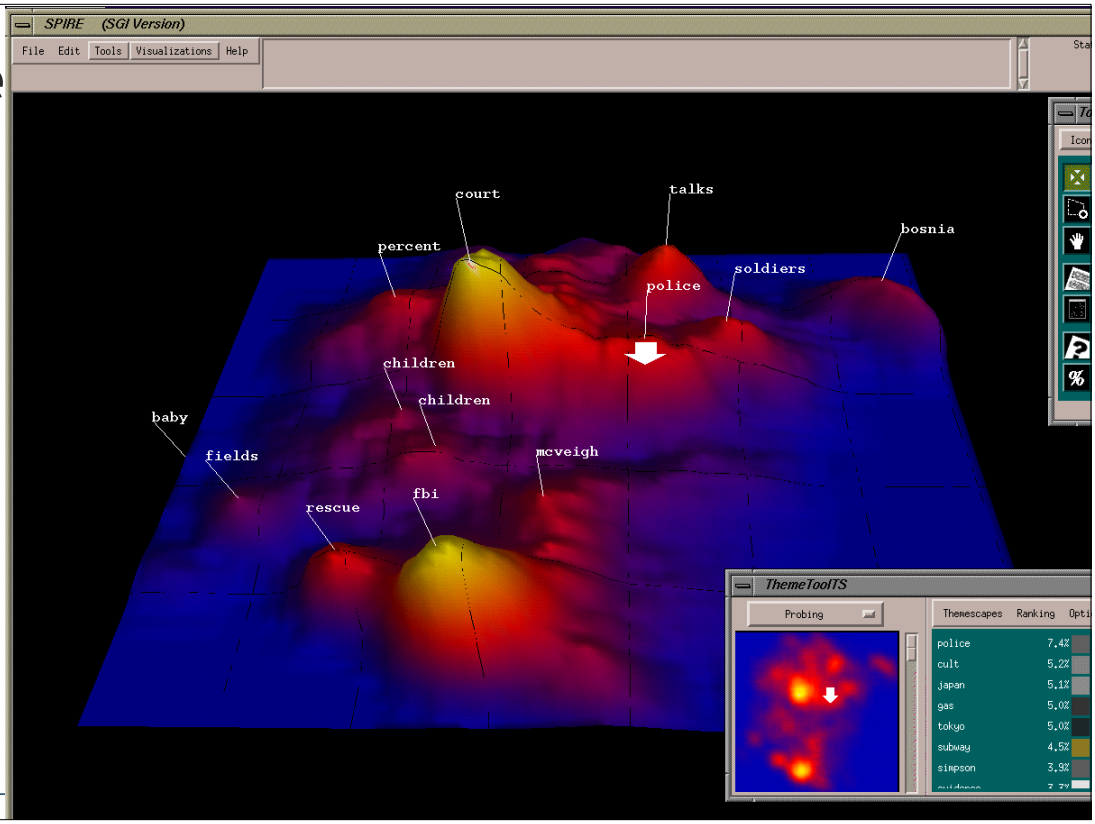
cluster of similar documents
close clusters are more similar to each other than clusters far away.
"Two level of similarity"



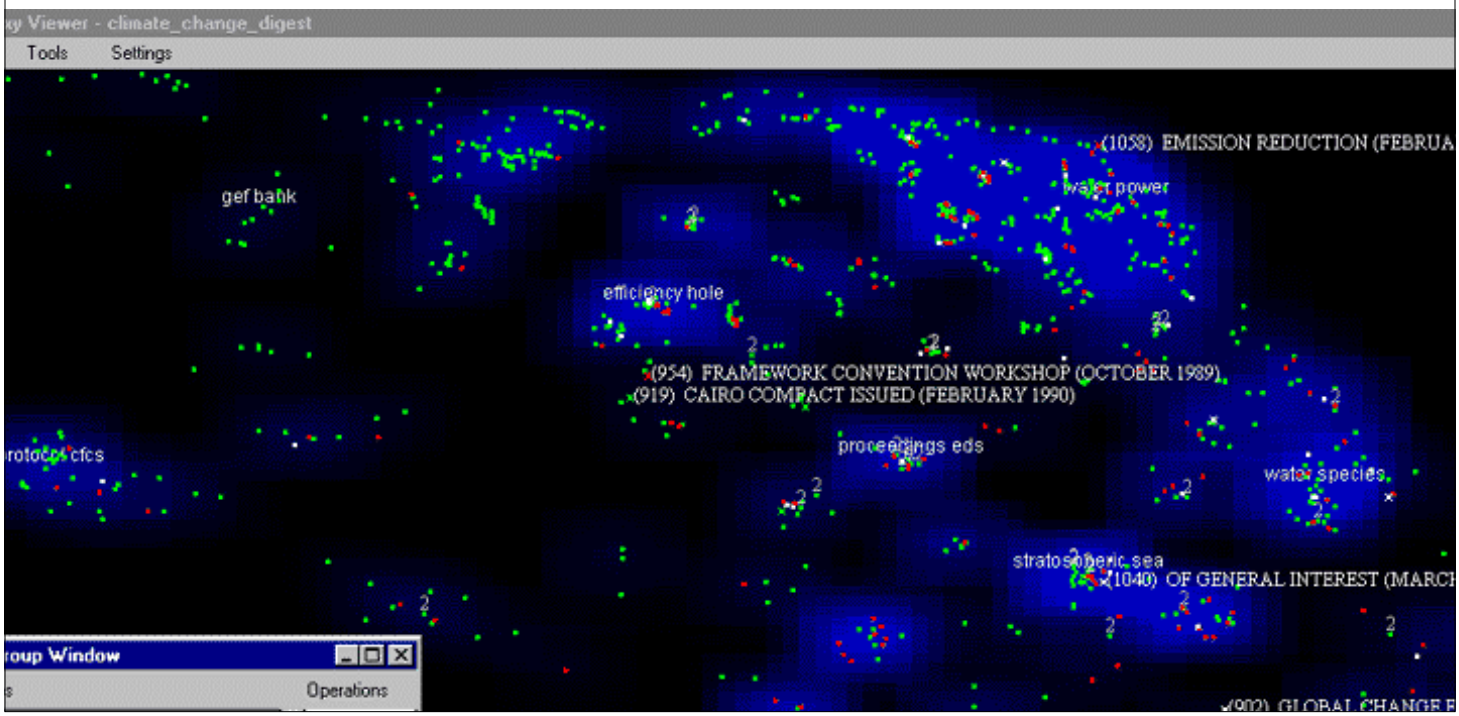
Themescapes

- 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



WebTheme

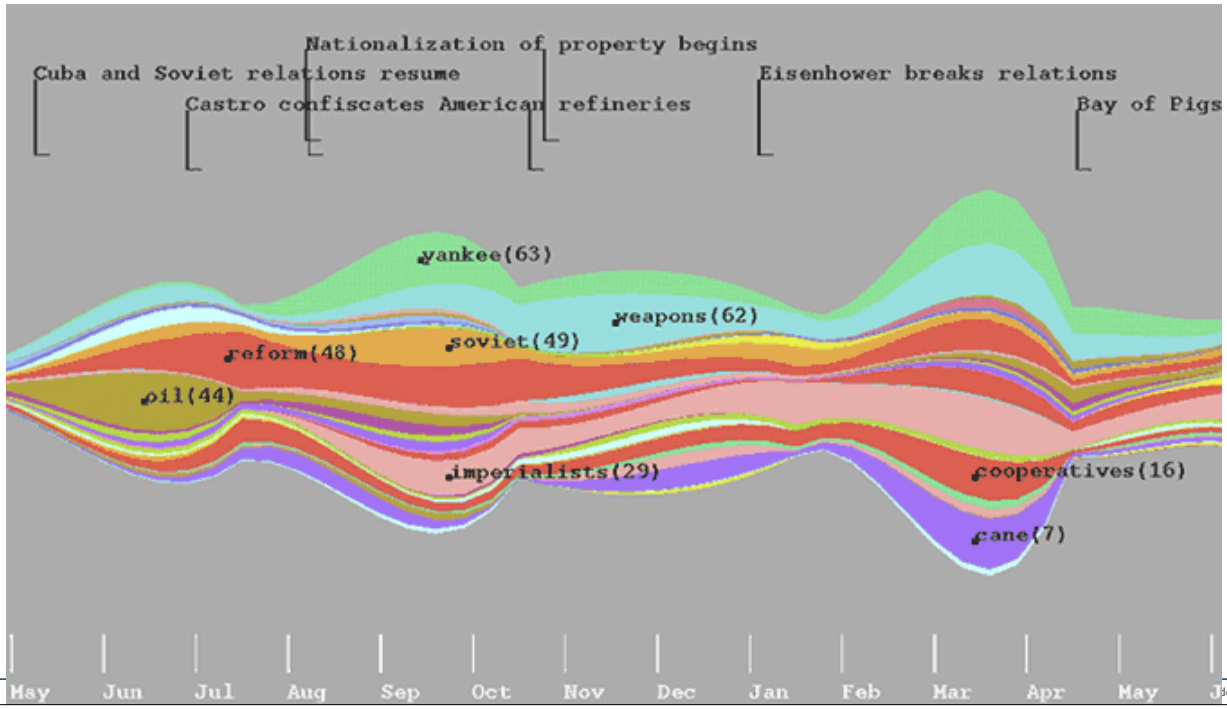


Temporal Issues

- 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

Topic Modeling

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

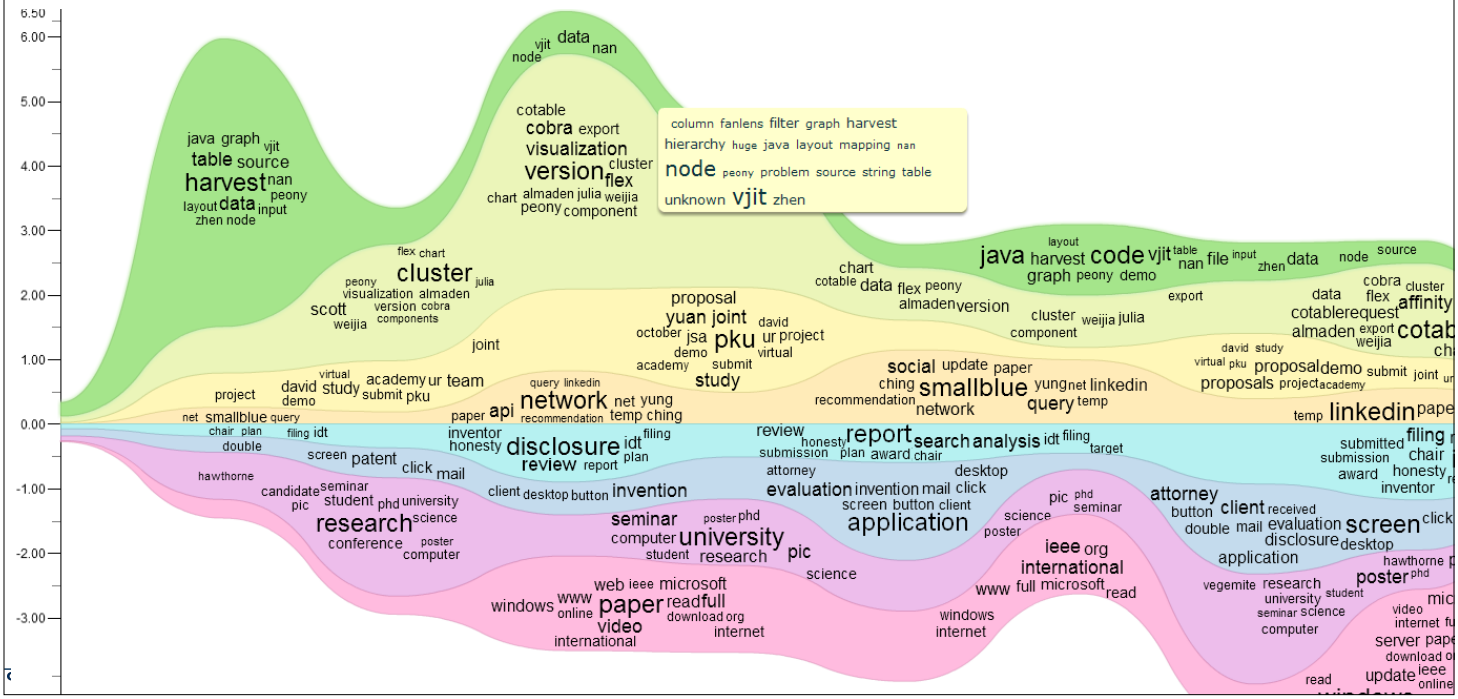
TIARA

- 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

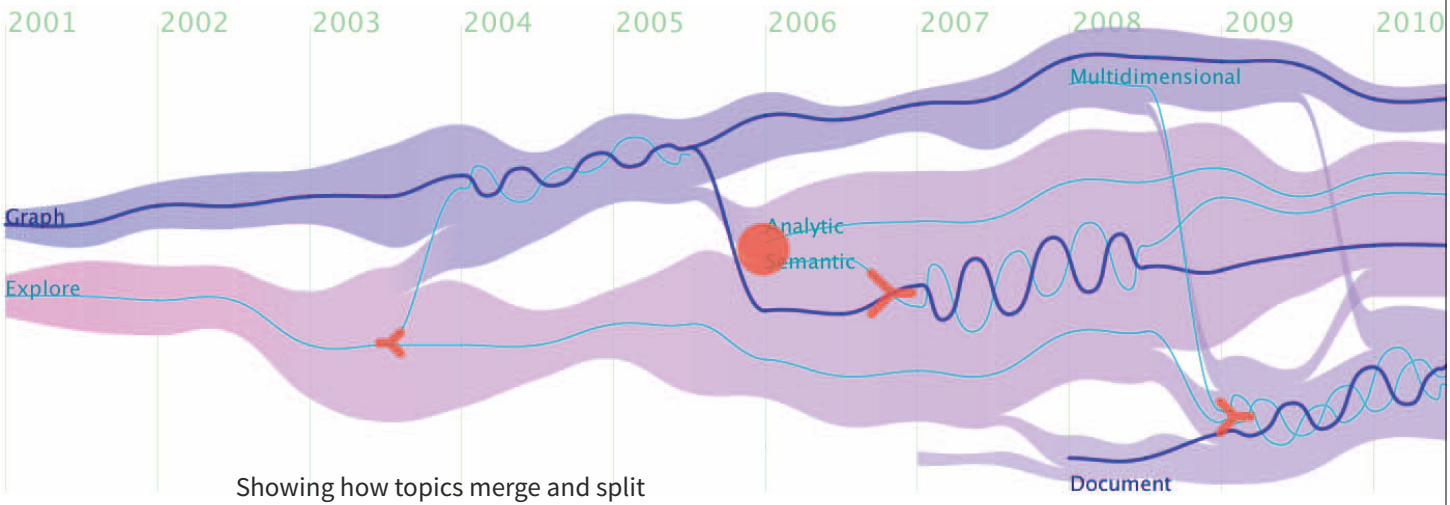
Liu et al
CIKM ‘09, KDD ‘10, VAST ‘10

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).



TextFlow



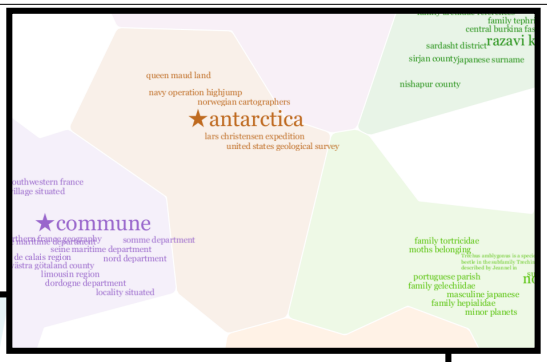
Showing how topics merge and split

Cui et al
TVCG (InfoVis) '11

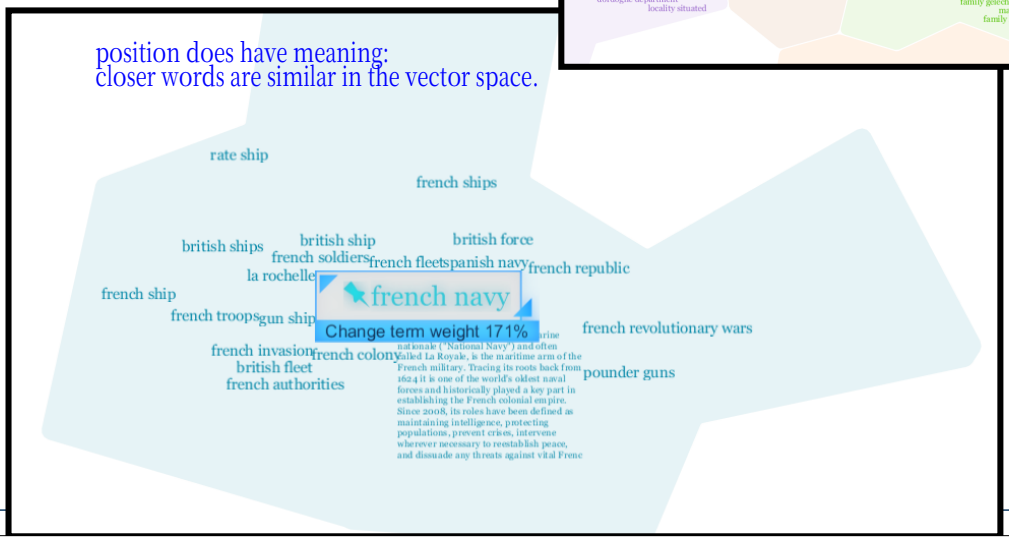
back to Word Clouds

- Recall: drawbacks of word clouds included:
  - position had no meaning
  - size was artificial at times (word length confounds size)
  - interaction is limited

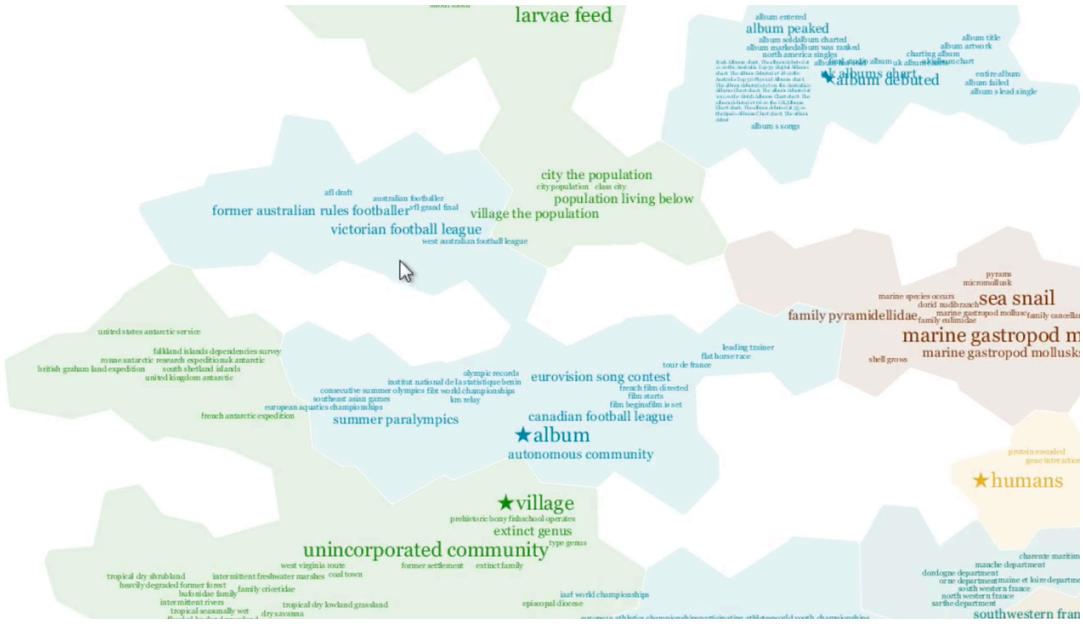
TexTonic



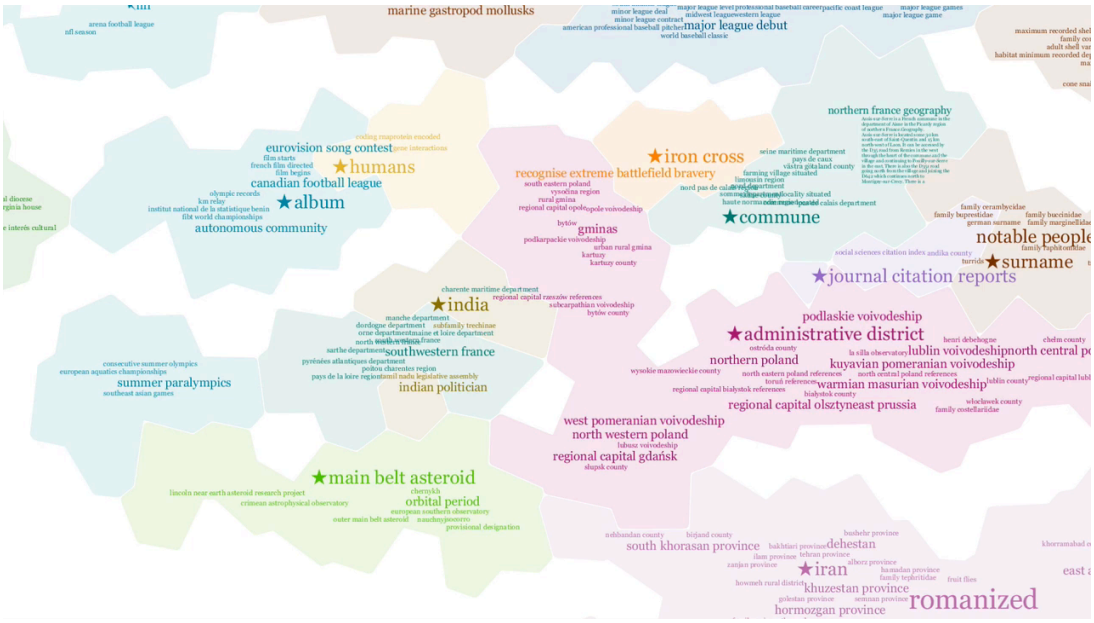
position does have meaning:
closer words are similar in the vector space.



TexTonic



TexTonic


marine gastropod mollusks
major league debut
northern france geography
eurovision song contest
★humans
canadian football league
★album
autonomous community
★iron cross
recognise extreme battlefield bravery
★commune
gminas
notable people
★surname
★journal citation reports
★india
southwestern france
indian politician
summer paralympics
podlaskie voivodeship
★administrative district
northern poland
kuyavian pomeranian voivodeship
warmian masurian voivodeship
regional capital olsztyneast prussia
west pomeranian voivodeship
north western poland
regional capital gdańsk
★main belt asteroid
orbital period
south khorasan province
dehestan
★iran
khuzestan province
hormozgan province
romanized