University of Stuttgart
Germany

Trends and Topics

How visual approaches foster synergetic effects by combining linguistic analyses and interactive exploration

Steffen Koch, Dennis Thom, Florian Heimerl, Harald Bosch, Han Qi, et al.

Workshop 'Data mining and its use and usability for linguistic analysis' | Saarbrücken | 2015-03-13

Institute for Visualization and Interactive Systems

Overview

- Information Visualization
- Visual (Text) Analytics
- Monitoring Twitter
- Understanding Scientific Communities
- Future Directions

Information Visualization (InfoVis)

- Is an algorithmic procedure that
  - turns abstract data
  - into a visual representation human users can interpret
- It is a translation process that bridges the gap between quantitative data and human perception and cognition
- Interaction is an integral part of Information Visualization
- Visualization as a research field is interdisciplinary

"The purpose of computing is insight not numbers" (Hamming 1962)

"The purpose of visualization is insight not pictures" (Shneiderman 1999)

Goals of Information Visualization

Visual Analytics

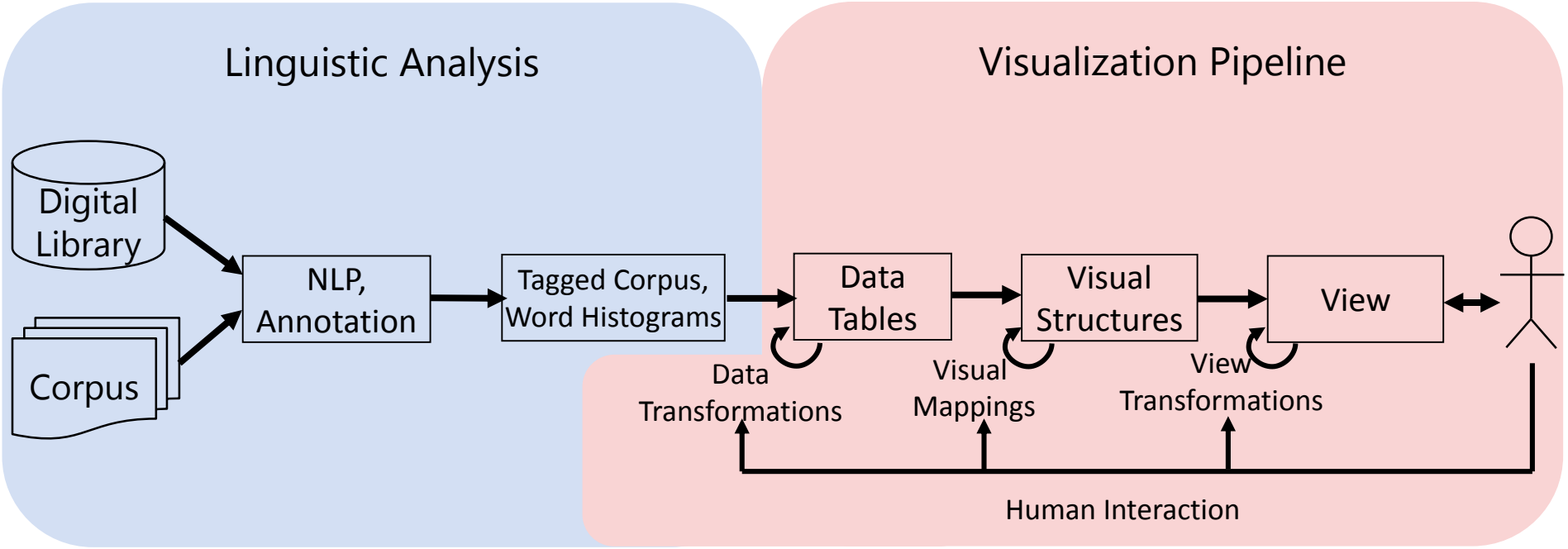
- Interactive, visual analysis
  - Visual information extraction
  - Explicate non-obvious relationships
  - Cope with uncertainties
  - Test hypotheses
  - Steering of ML methods

- Presentation/communication
  - Make complex information accessible
  - Give orientation
- Interactive exploration
  - Filtering
  - Visual search

Information Visualization

Visual Text Analytics

- (Computer) Linguistic analysis
- Decoupled from visualization steps
- No interactive control/selection of linguistic annotation/processing

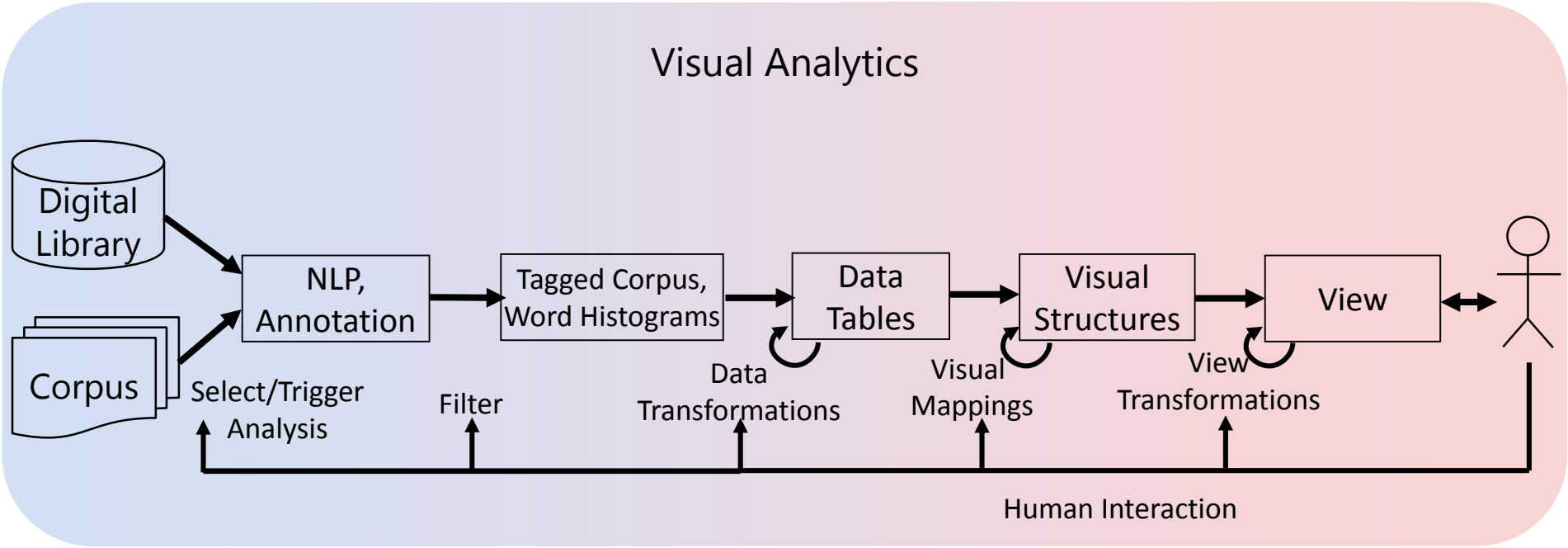


[adapted from Card et al. 1999]

Visual Text Analytics

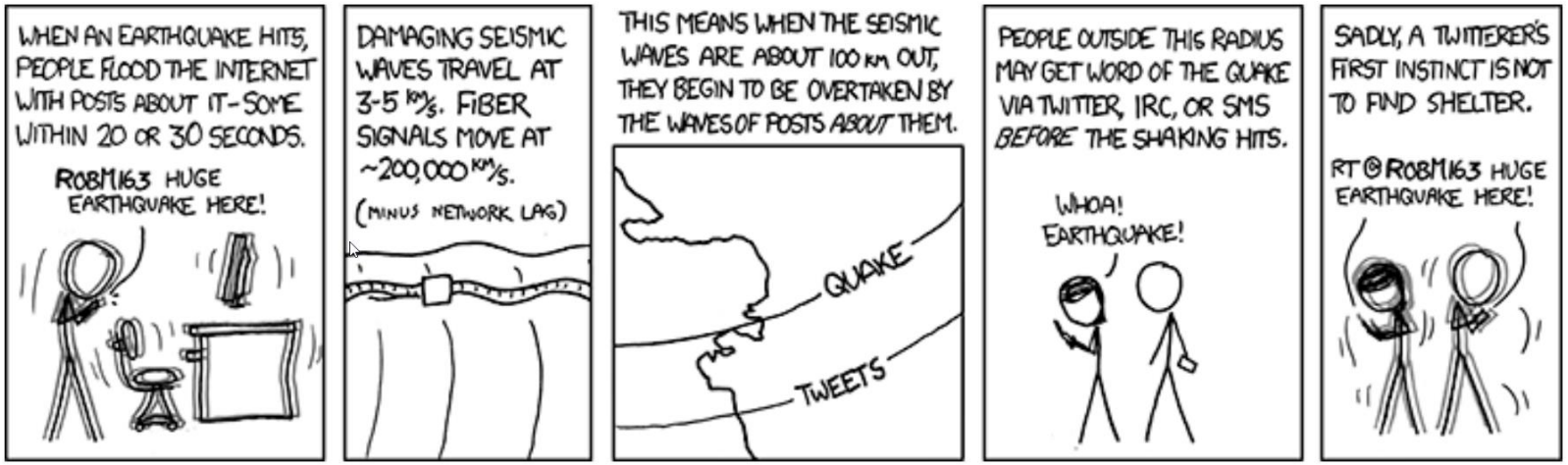
- Integration into Visual Analytics system

Interaction provides the glue for this coupling



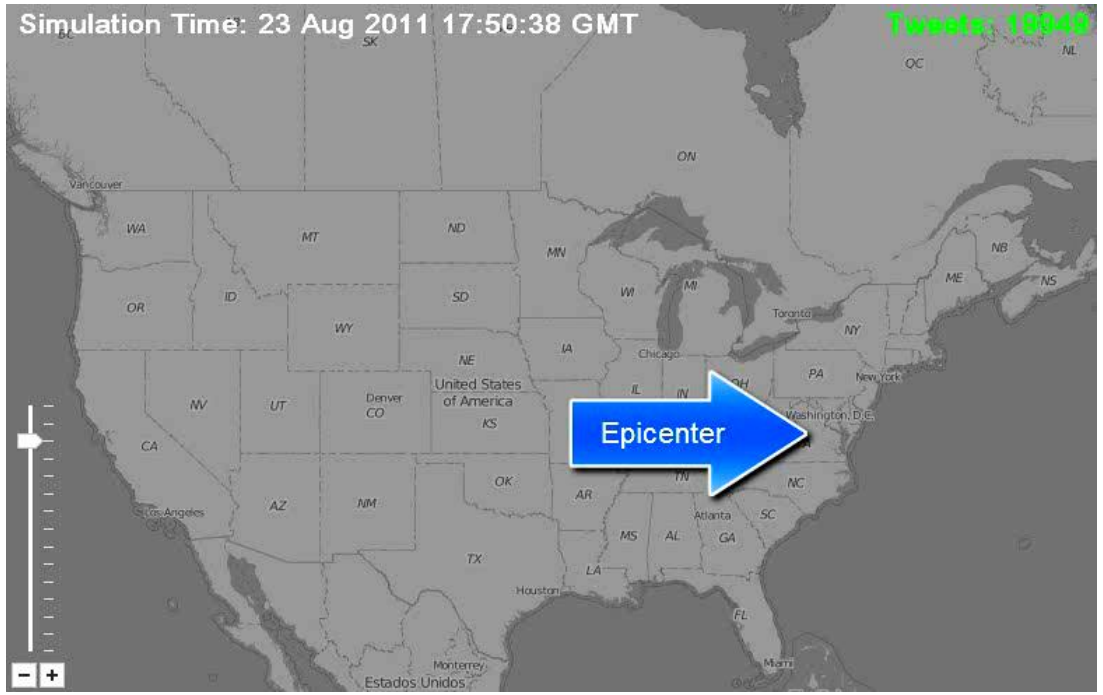
[adapted from Card et al. 1999]

Analyzing MicroBlogs
Motivation – The Speed of Tweet



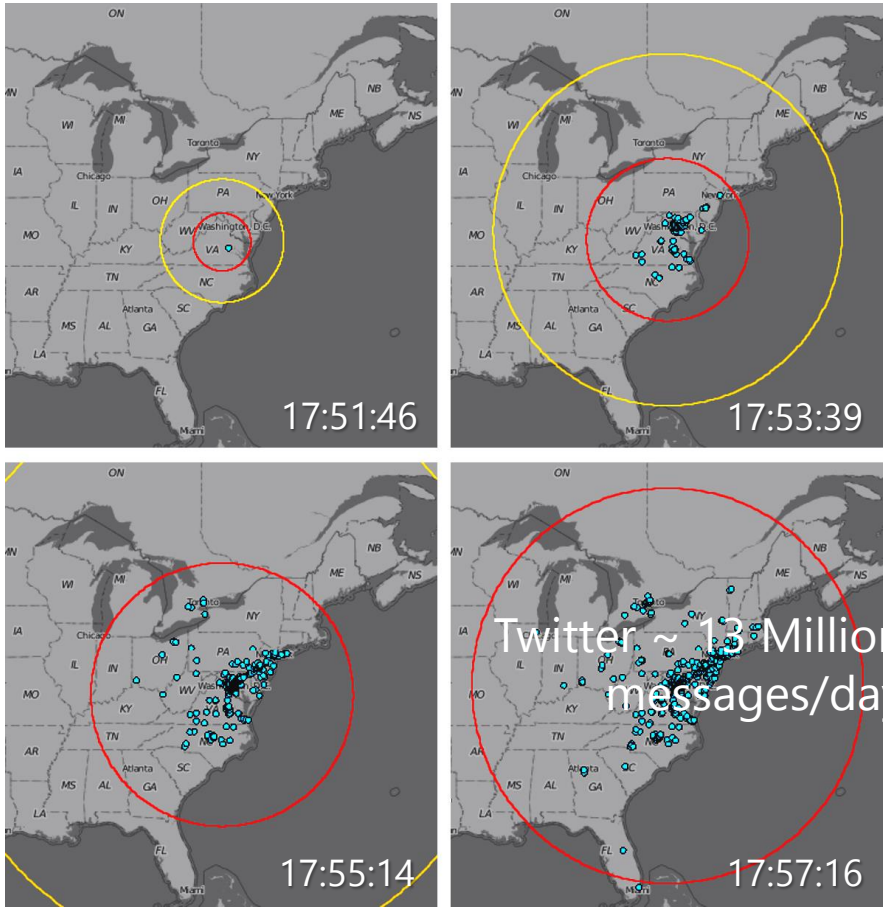
Motivation – The Speed of Tweet

- Visualization of Virginia Earthquake 2011
  - Yellow = P-Wave
  - Red = S-Wave
  - Blue = Tweets
- Event demonstrates the high timeliness and distribution of social media reactions



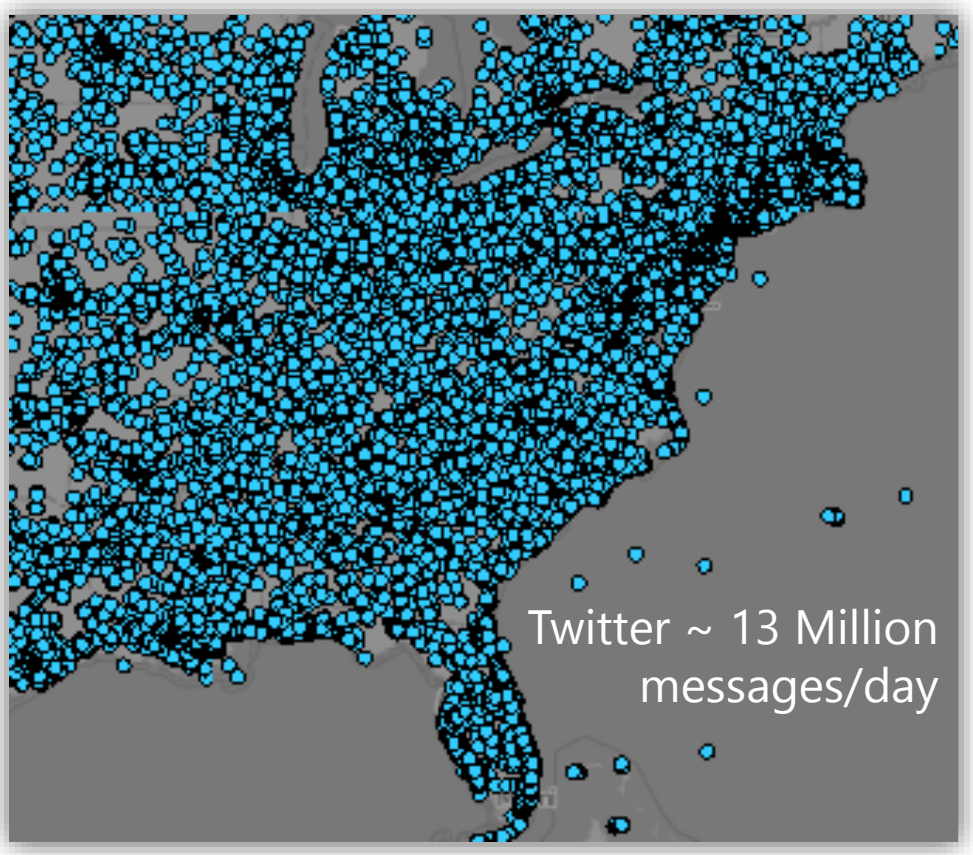
Motivation – Situation Awareness

- Information extracted from the data could be of great value for decision makers in:
  - Disaster Management
  - Public Safety
  - Disease Control
- How to identify relevant information and produce meaningful situation overviews from millions of messages in real-time?



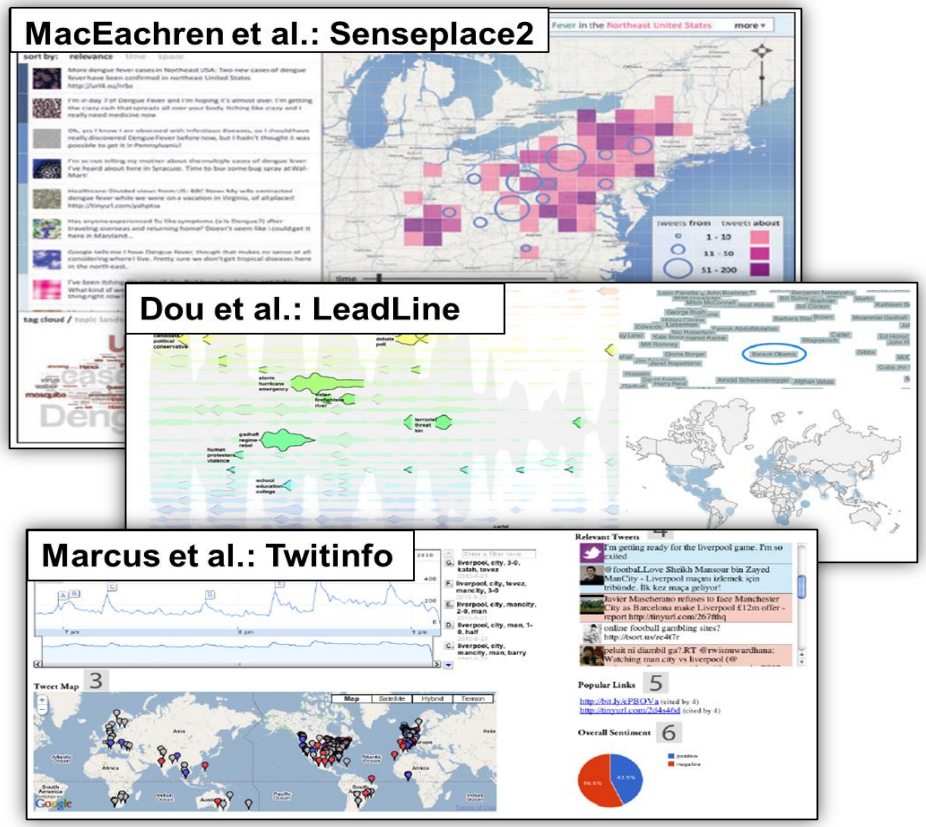
Motivation – Situation Awareness

- Information extracted from the data could be of great value for decision makers in:
  - Disaster Management
  - Public Safety
  - Disease Control
- How to identify relevant information and produce meaningful situation overviews from millions of messages in real-time?



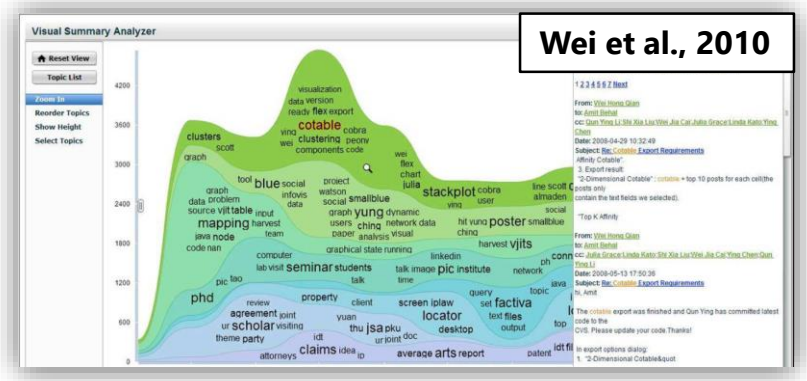
Visual Analytics

- Visual Analytics (VA) integrates data visualization with statistical models to solve such “big data” challenges
- Previous approaches have primarily employed text aggregation models to support visual exploration and overviews of social media
- However, so far the systems are not applicable in context of real-time decision-making, due to:
  - Limited scalability
  - Only qantitative analysis
  - Rate limits / query costs



Visual Analytics – Machine Learning Components

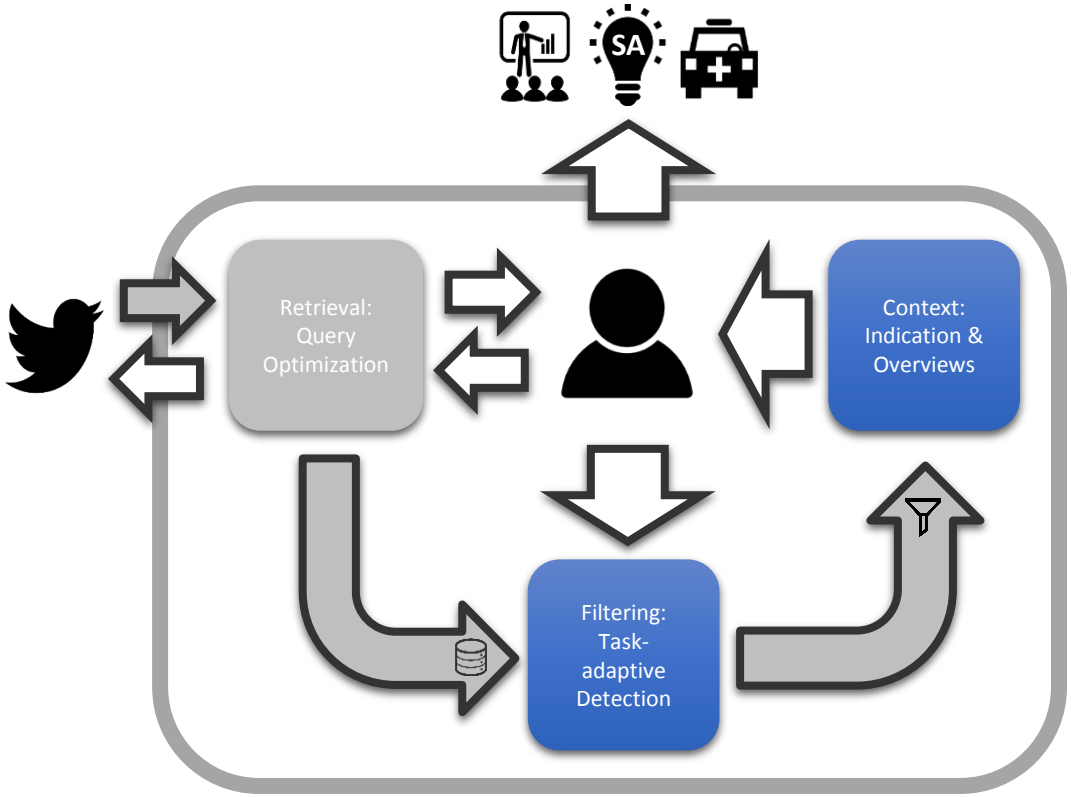
- Thesis highlights that statistical models play two central roles in visual analytics:
  - **Overview & Discovery**: Unsupervised methods are used to organize and aggregate data for interactive exploration
  - **Filtering & Detection**: Visual interaction with supervised methods is used to optimize recognition and to understand model behavior
- Techniques have not been used in combination in order to quickly adapt to changing situations



Social Media Analysis Cycle

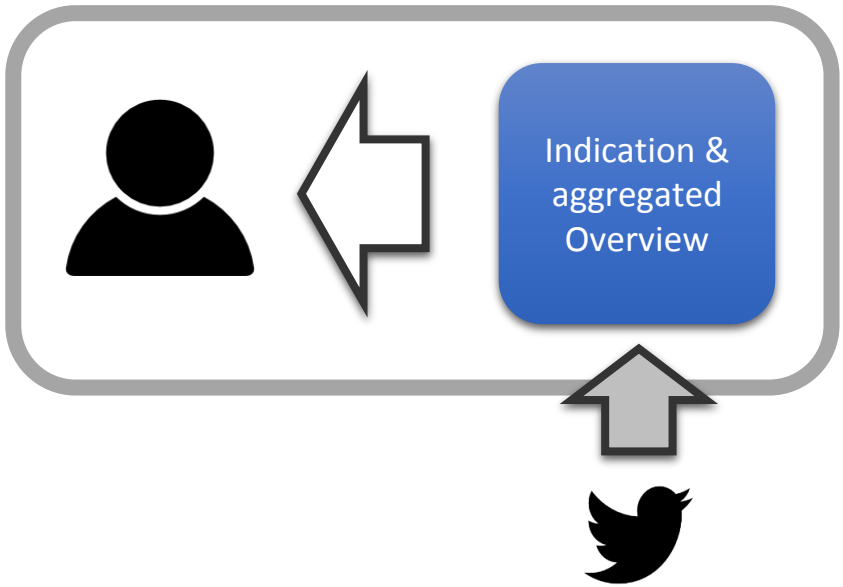
- Created an analysis loop that connects supervised with unsupervised methods to allow immediate transition between discovery and detection

- Incorporates novel techniques to support three stages of analysis:
  - Data Retrieval
  - Data Filtering
  - Data Contextualization



Indication and Overview

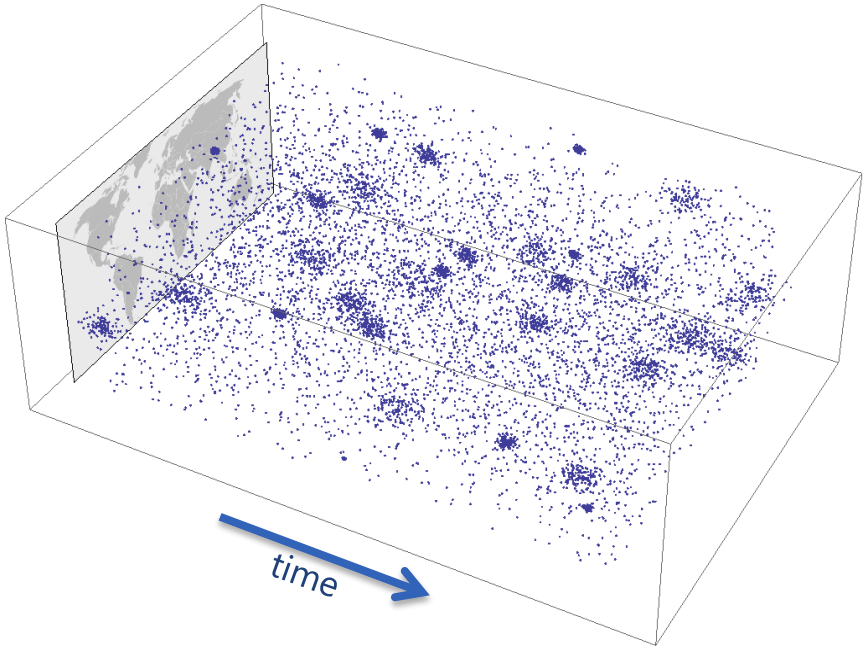
- The nature of an ongoing situation is unclear or it is even unknown that something relevant has happened



GREY = Data Flow
WHITE = Information Flow

Spatiotemporal Anomalies

- Assumption 1: Event is the most central entity in situation awareness
- Assumption 2: Events generate spatiotemporal clusters of similar tweets

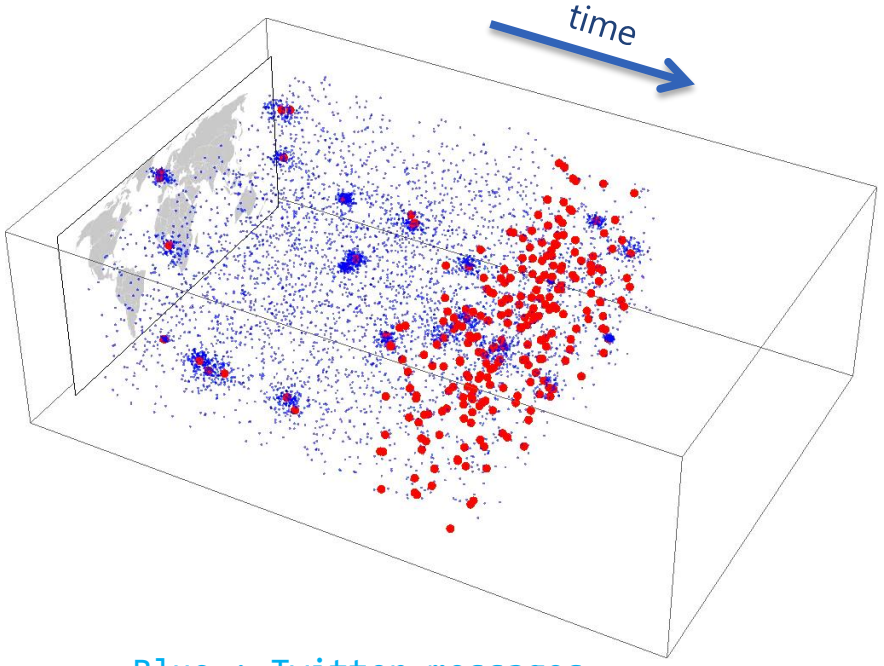


Spatiotemporal Anomalies

- Traditionally, analysts would enter keyword queries to find this kind of clusters
- Idea: Revert the process by detecting such clusters in the data and use them to generate a visual overview of what should be searched for

Anomaly Discovery

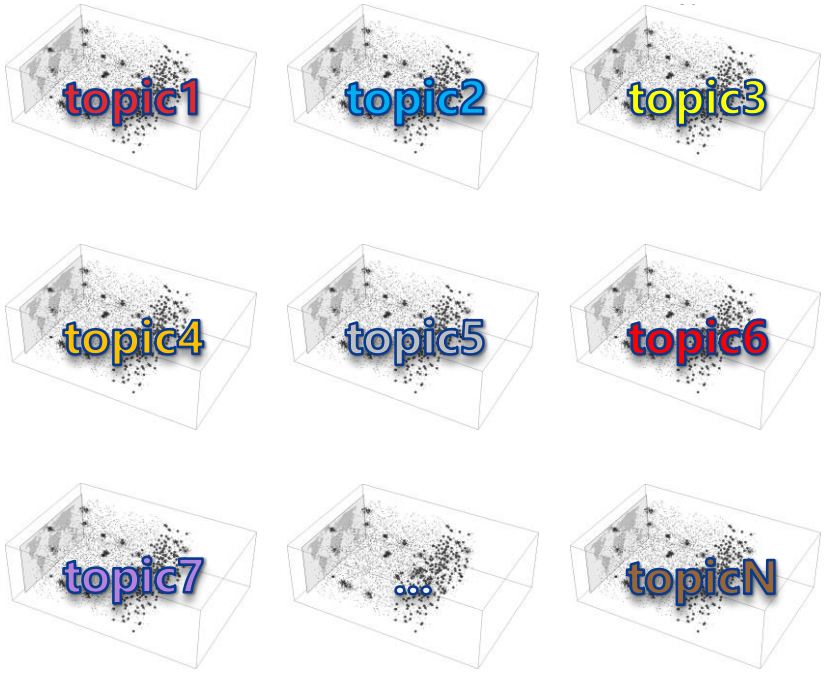
- Streaming-enabled cluster analysis based on K-Means:
  - Instead of a fixed number of centroids (means), a splitting mechanism is employed
  - A sliding window is used to evaluate/discard clusters once they turn stale



Blue : Twitter messages
Red : Centroids

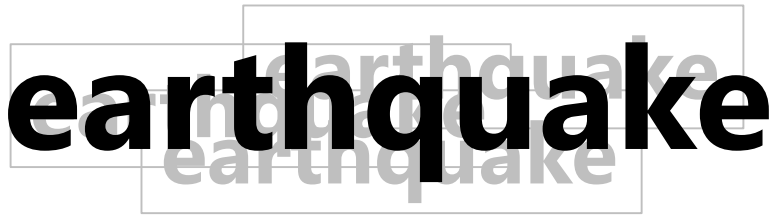
Anomaly Discovery

- Clustering performed seperately for any observed topic to account for „clusters within clusters“



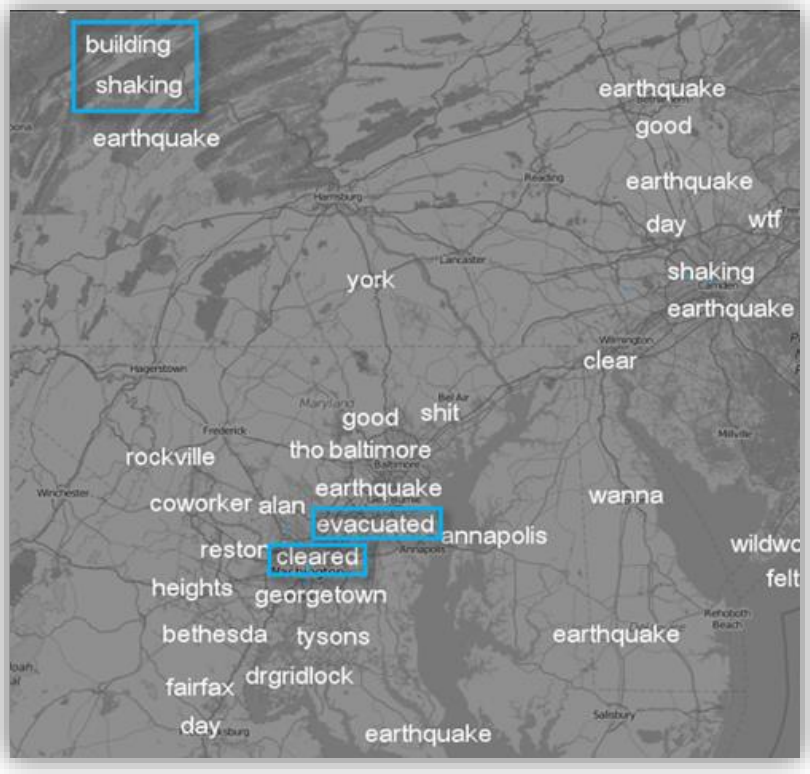
- Each new message is assigned to topics and only the corresponding cluster branches are updated

Anomaly Visualization



- Detected clusters are placed as labels on the map – collision resolution produces tagcloud layout
- Similar overlapping labels are aggregated to counteract overfitting and allow adaptive semantic zoom

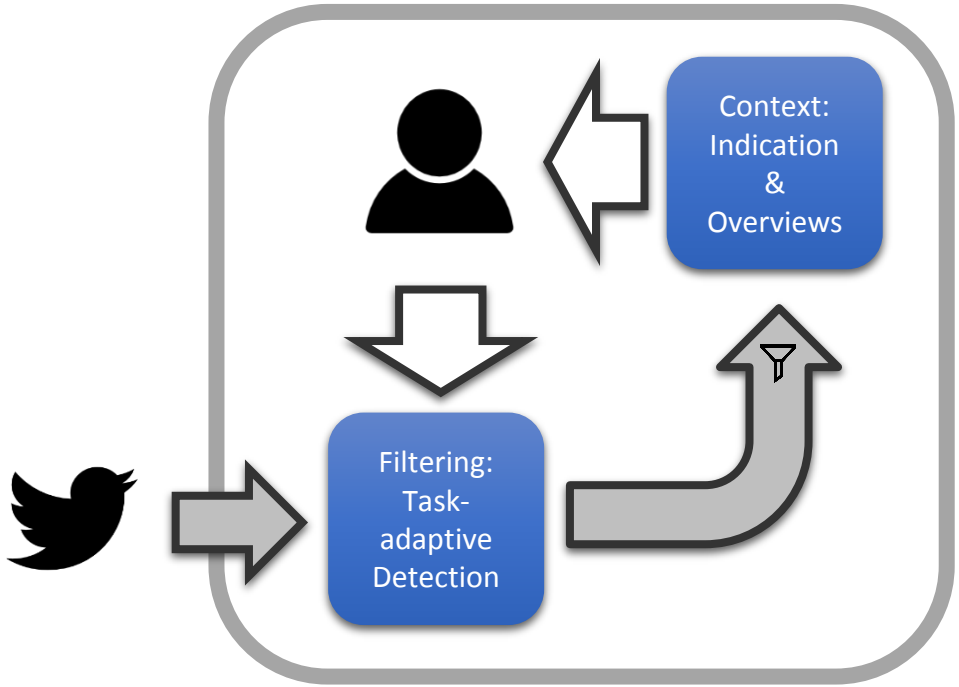
Anomaly Visualization



- Finally, investigation of high zoom-levels reveals individual sub-events connected to the larger event

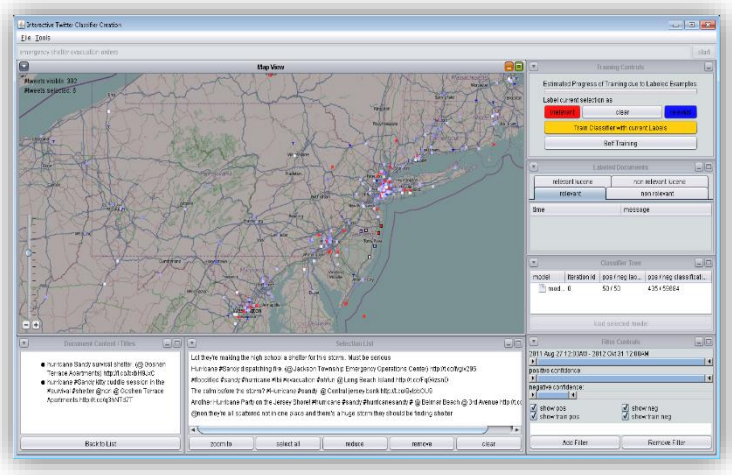
Task-adaptive Filtering & Detection

- Following initial assessment, analysts in situation awareness usually require high recall

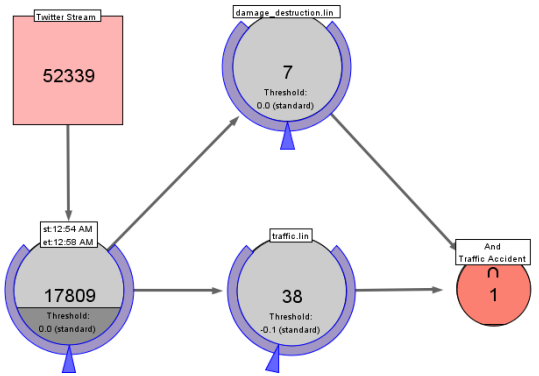


Classifier creation/configuration/orchestration

- Create linear binary classifiers (e.g. SVM) in a visual, exploratory fashion based on recorded data from past events



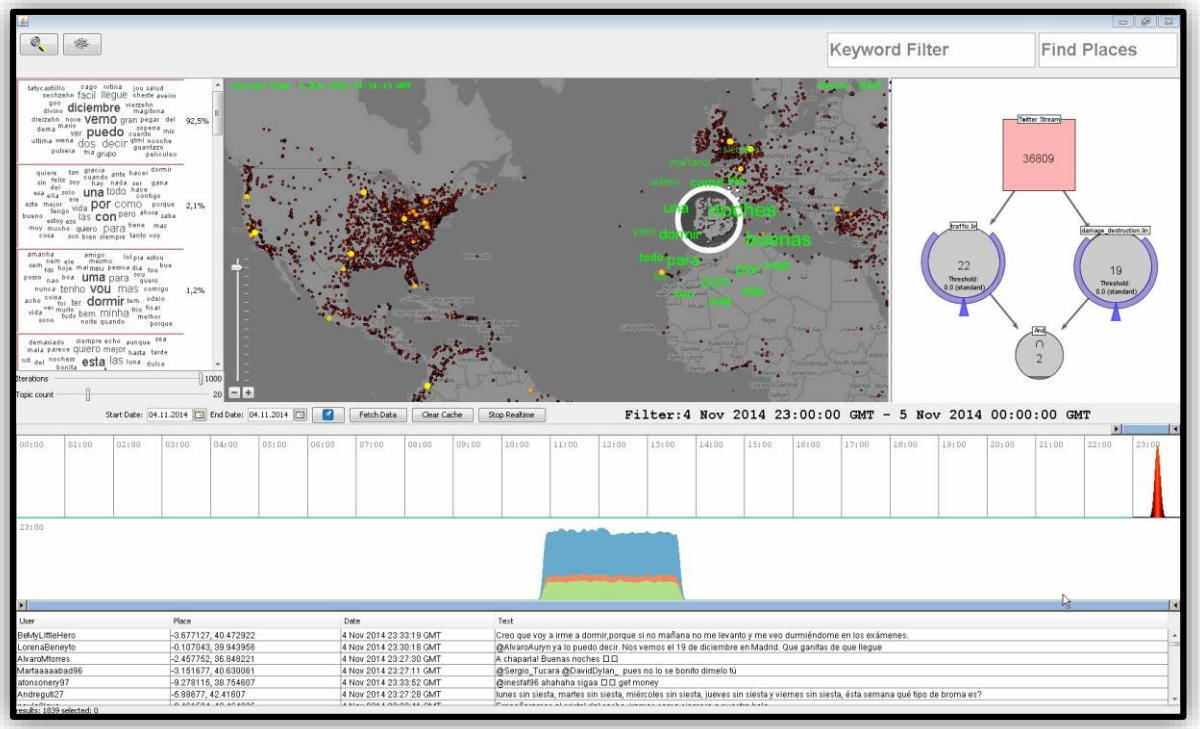
- Interactive filter/flow-metaphor allows classifier orchestration and configuration during ongoing monitoring



ScatterBlogs: Integrated analysis environment

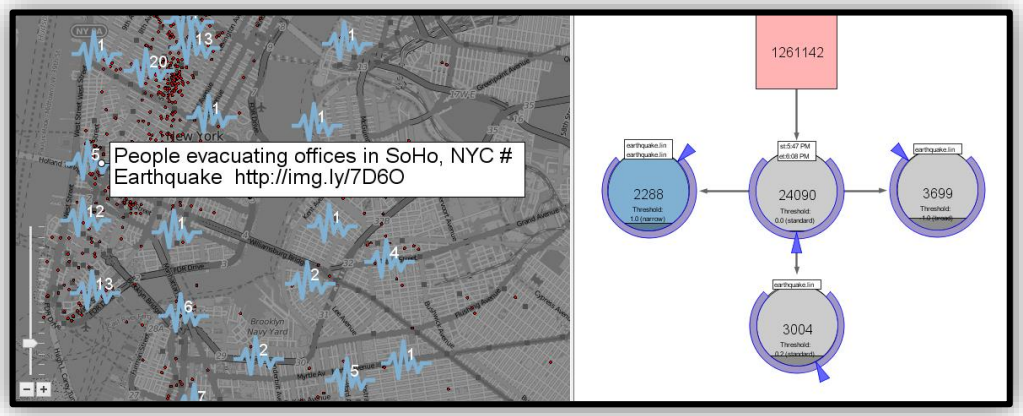
- Components implemented as part of the ScatterBlogs VA system

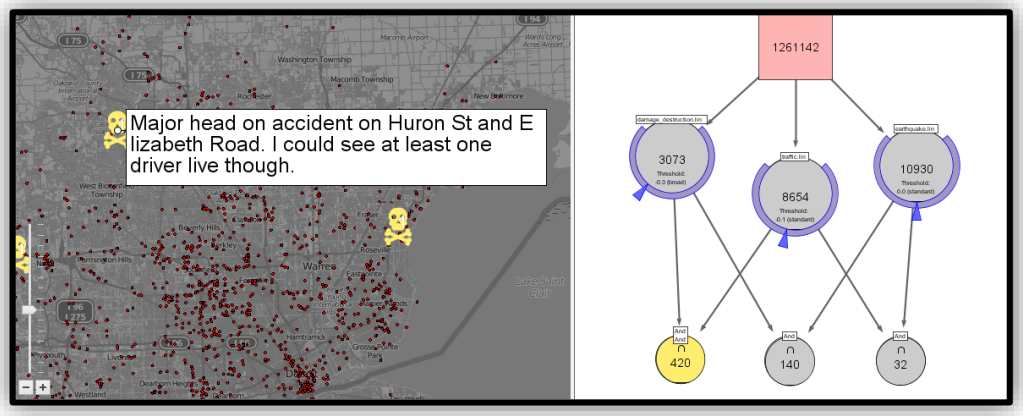
- Additionally provides basic search & filtering as well as real-time message visualization and exploration



Earthquake Case Study

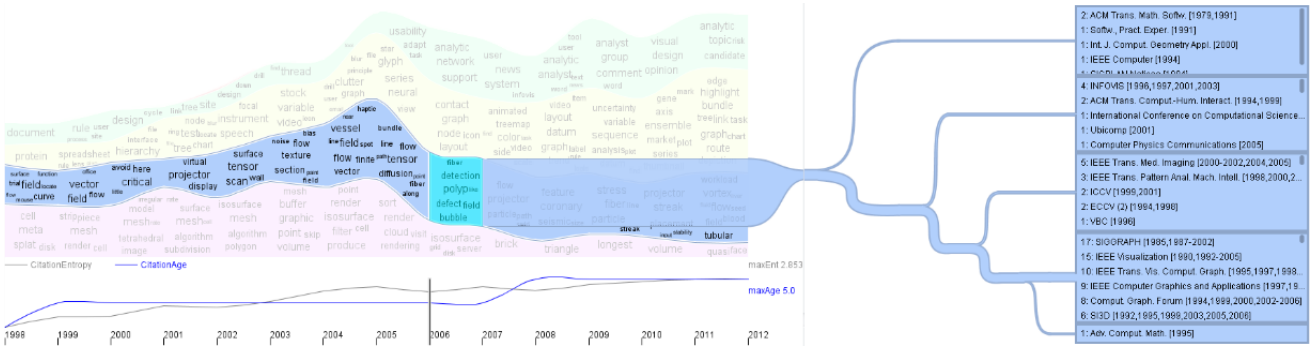
- Applied classifier orchestration to 2011 Virginia Earthquake
- Classifier refinement can be used to fine-tune detection rates and compare different configurations
- By Boolean combination of classifiers, analysts can create ad-hoc detection structures





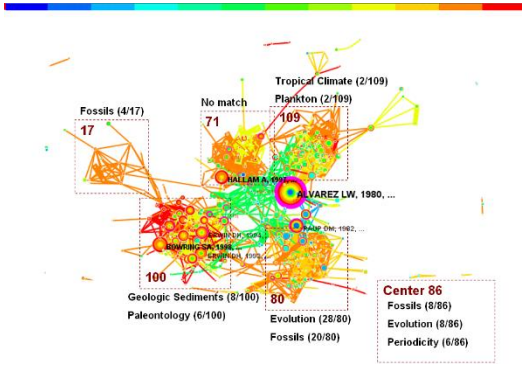
Next Example: Analyzing development of scientific communities

- Two use cases/data sets
  - Scientific literature analysis
  - Patent analysis
- In previous work
  - Mainly structures of citation networks (often node links diagrams)
  - No combination with content information
- Goals
  - Support retrieval
  - Understand developments, topics, trends in communities

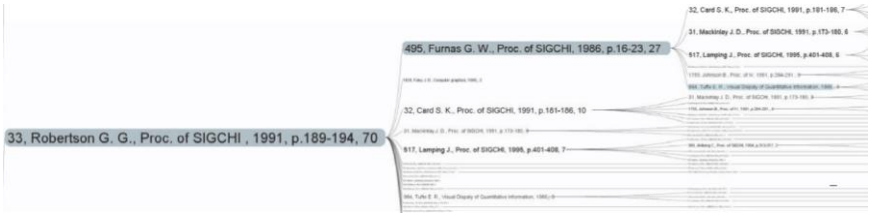


Related Work

- Chen, 2006: CiteSpace II

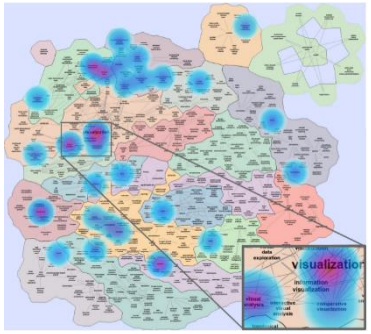


- Zhang et al., 2009: Visualizing the Intellectual Structure with Paper-Reference Matrices



- Lee et al., 2004: PaperLens

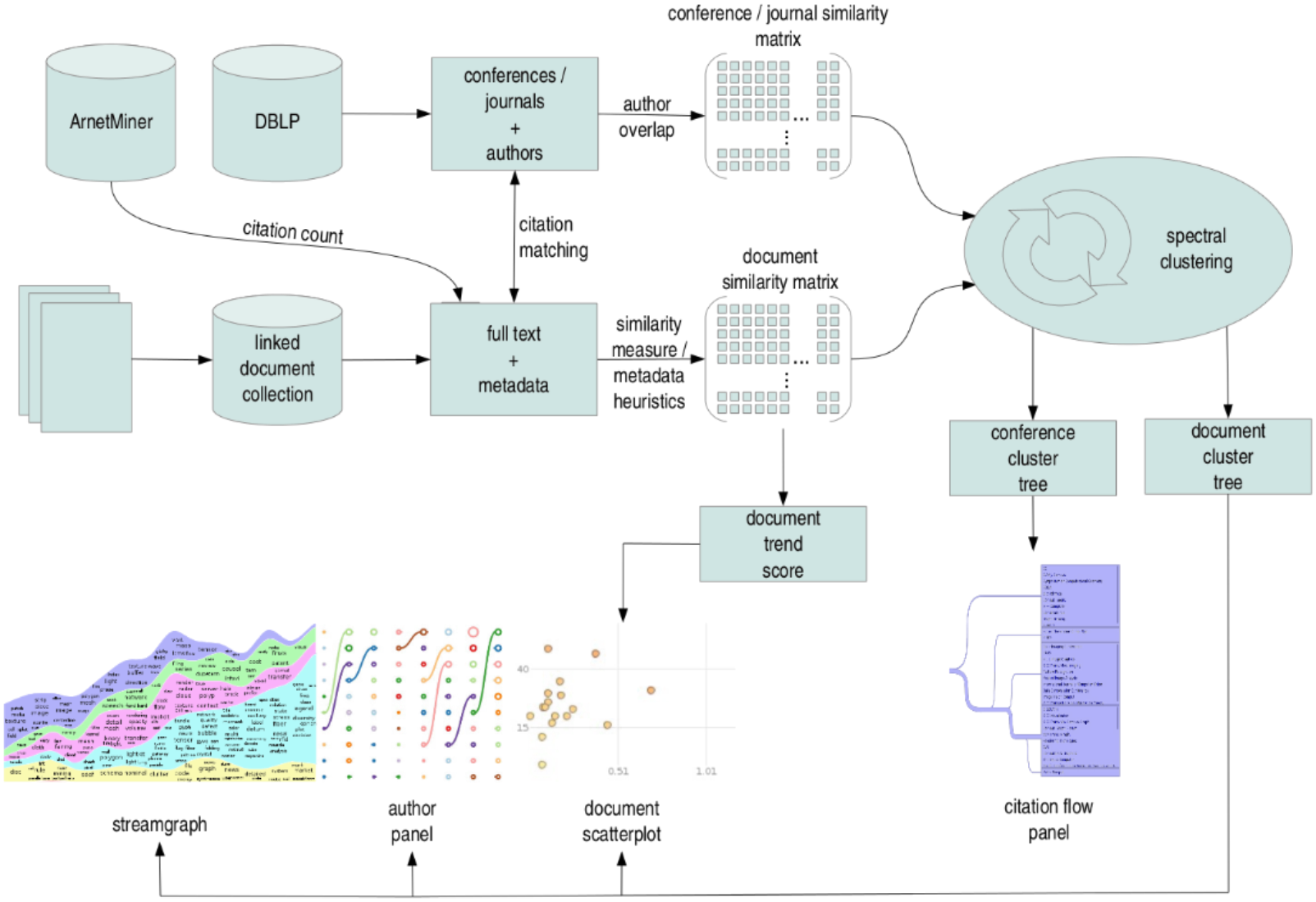
- Fried & Kobourov, 2014: Maps of Computer Science



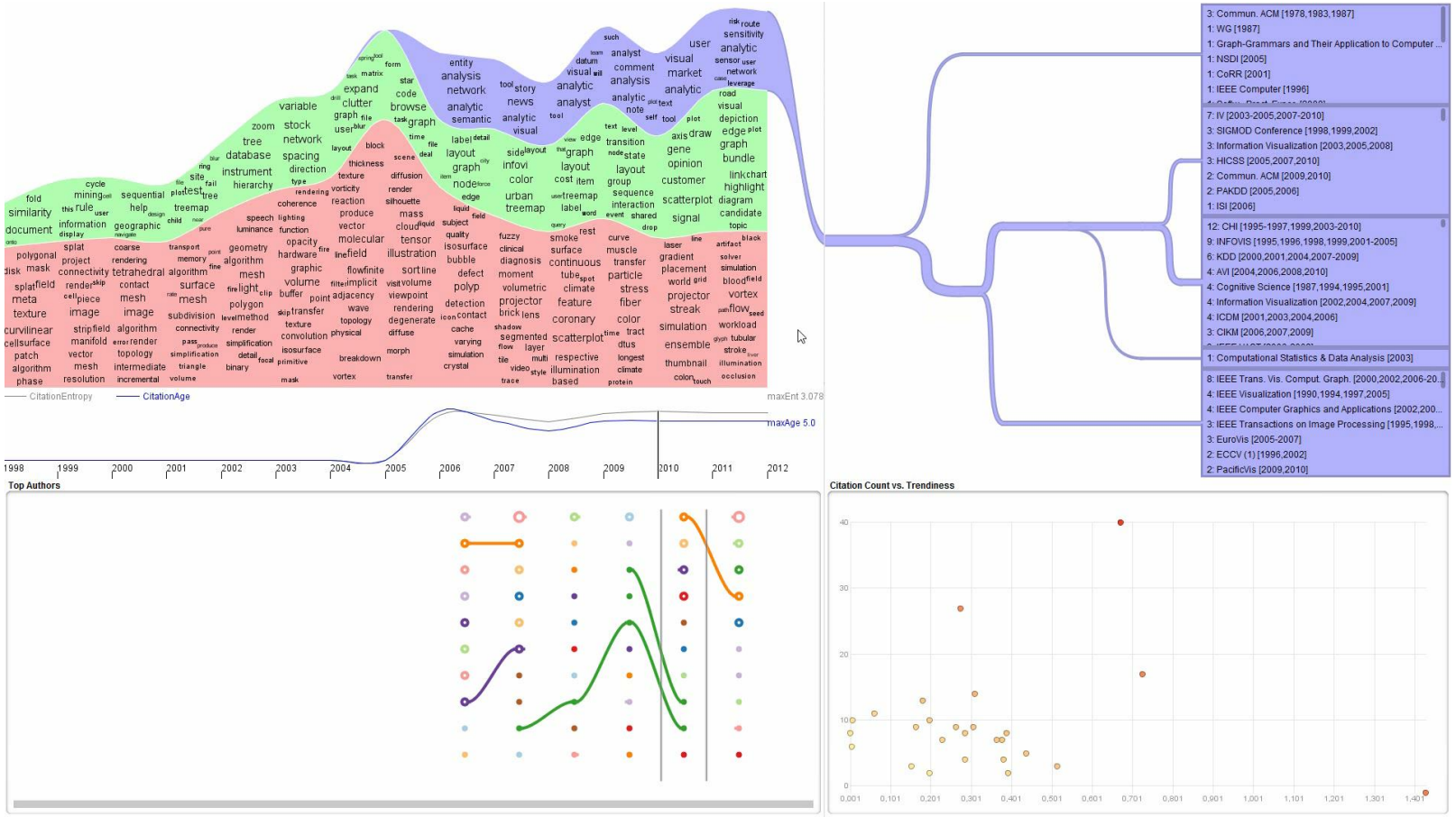
Example Dataset

- VIS dataset
- Extracted from Vis(Week) publication pdfs (1998-2012)
  - Omnipage for OCR
  - ParsCit [Councill et al.] for citation extraction
  - Manual cleaning of metadata
    - Authors, title, abstract
- Extracted citation strings
  - Structured output from ParsCit
  - Map titles to DBLP [Ley]
    - Using Levenshtein Algorithm
  - Use unambiguous conference / journal id from DBLP
  - Successful for ~70% of all references

Backend

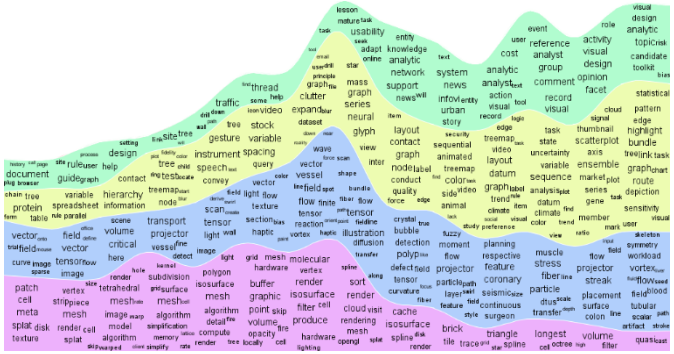

conference / journal similarity matrix
ArnetMiner
DBLP
conferences / journals + authors
author overlap
citation count
citation matching
document similarity matrix
spectral clustering
linked document collection
full text + metadata
similarity measure / metadata heuristics
conference cluster tree
document cluster tree
document trend score
streamgraph
author panel
document scatterplot
citation flow panel

System Overview



Document Grouping and Clustering

- Documents can be group according to metadata
  - e.g. conferences,
  - authors, or
  - keywords
- Clustering of documents
  - e.g. document content, or
  - citations
- Document similarity matrix for clustering
  - Different similarity measures

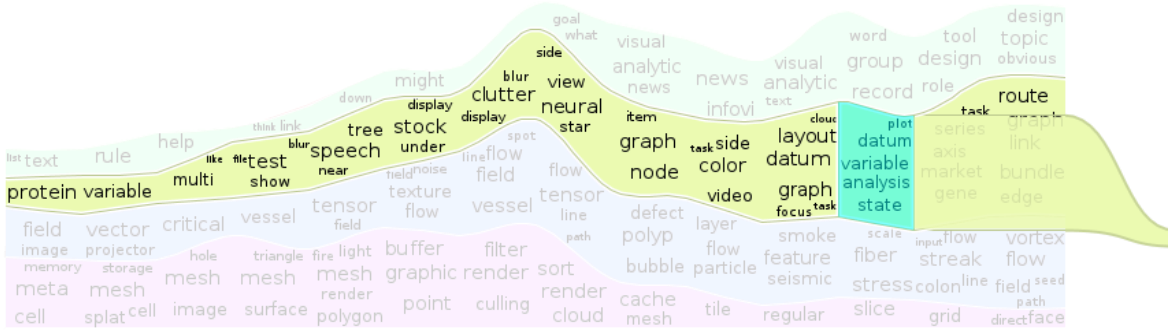


Document Similarity and Clustering

- We created two document clusterings
  - Content: using cosine similarity
  - Citations: using Jaccard coefficient
- Spectral Clustering
  - Similarity matrix contains similarity graph of documents
  - Recursively partition the graph (hierarchical)
  - Criterion for cut size
    - small volume of severed edges, and
    - balanced partition sizes
  - We use Normalized Cheeger Cut [Bühler & Hein]
    - $NCC(C, \overline{C}) = \frac{Cut(C,\overline{C})}{\min(vol(C),vol(\overline{C}))}$
    - $Cut(C, \overline{C}) = \sum_{i \in C, j \in \overline{C}} w_{ij}$ and $vol(C) = \sum_{i \in C, j \in C} w_{ij}$

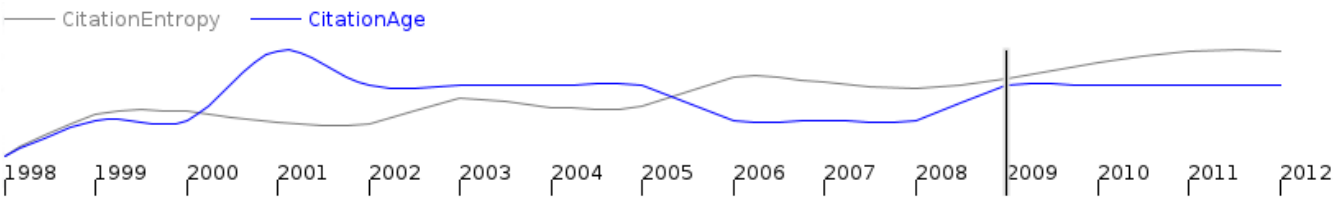
Keyword Extraction

- Wordclouds give information about clusters
  - Streams are split up into blocks for which clouds are created
  - Keywords are indicative of stream and block
- Keyword generation on paper abstracts
  - Tokenization, lemmatization
  - Stop word removal
  - G2 rating [Rayson & Garside]: difference in probability of occurrence in two bodies of text
- Keywords are placed in the block along a spiral path
  (same idea as in previous example)



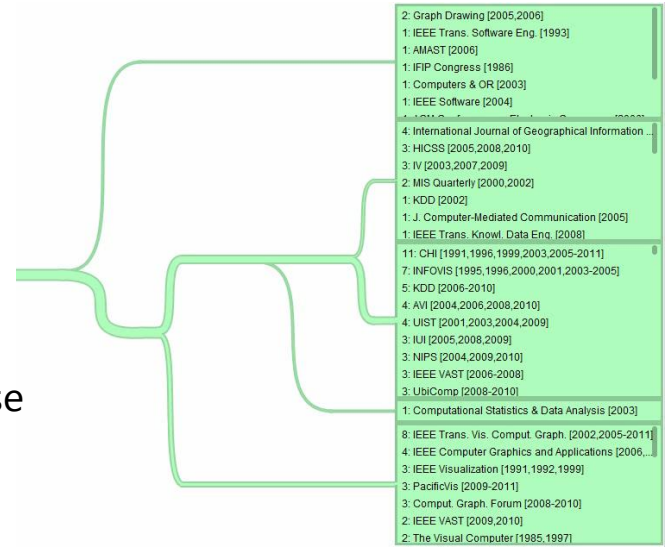
Citation Age and Citation Entropy

- Characteristic values of the documents in the clusters
  - Citation age
    - The average age of the citations in the reference list of the document
    - Gives information about how old cited material is
  - Citation entropy
    - The entropy with respect to the cited publication venues
      - $H = -\sum_{V \in Cited\_Venues} \frac{num\_cites(V)}{num\_cites} \log \frac{num\_cites(V)}{num\_cites}$
    - Gives information about the diversity of the cited material



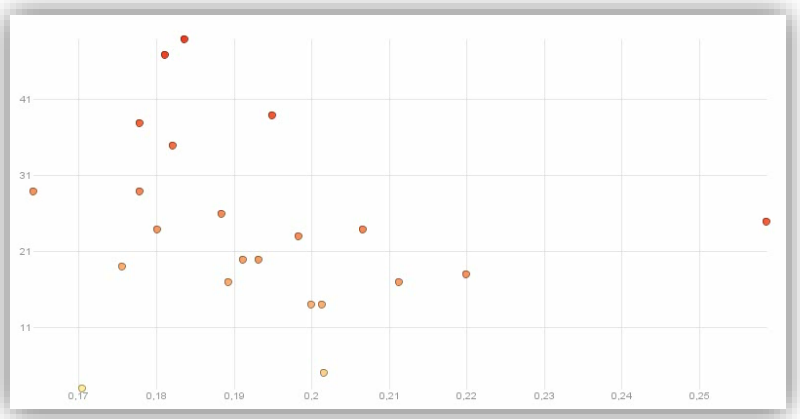
Clustering Publication Venues

- Goal: find a useful aggregation of conferences and journals that
  - Corresponds with user expectations about similar venues
  - Capture scientific communities
- We use publication venues from the DBLP dataset
  - Entries in the DBLP associated with publication venue
  - For each venue collect publications
  - Collect all author names from these publications
- Distance measure for conferences and journals
  - Jaccard coefficient as similarity measure
  - Persons with same name by coincidence treated as noise
- Clustering based on the spectral clustering algorithm
  - Hierarchical clustering
  - Users can change interactively change number of clusters



Trendiness Score

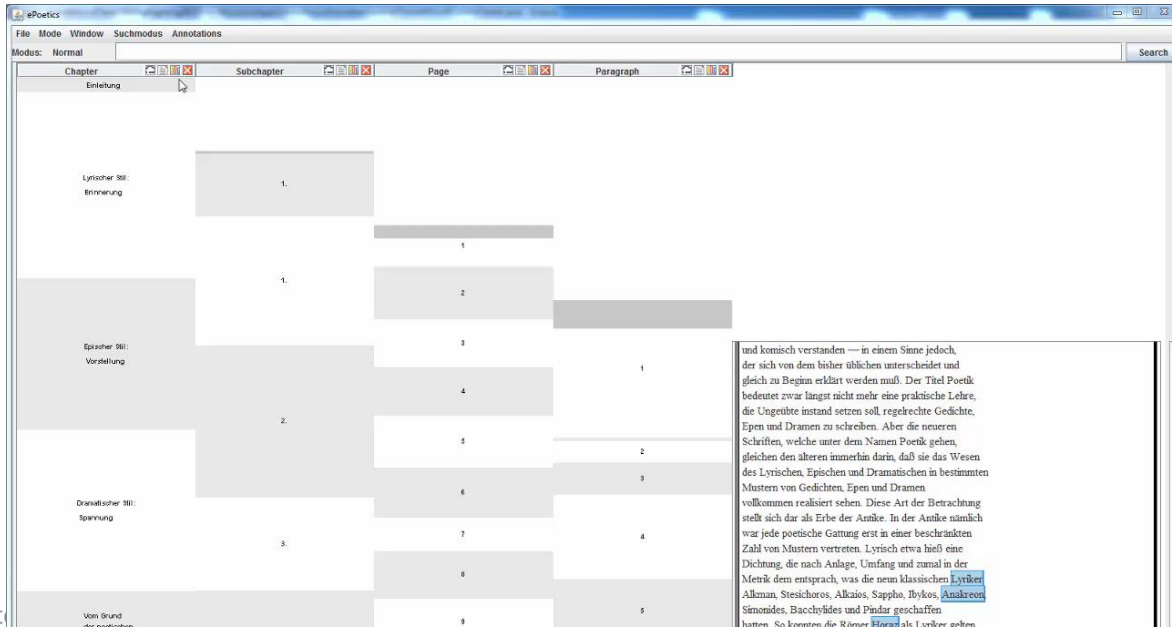
- Capture the trendiness of each publication
  - Documents that have trendy content will
    - Have little similar documents in the past
    - Have more similar documents in the future



- Trendiness Score
  - Idea: measure influence of previous papers to current paper and currents paper influence on subsequent papers based on the contents
  - We use a Gaussian kernel with size τ (estimated) to weigh similar papers (identified by cosine distance s_{ij})
    - $I_p = \sum_{j \in D_p} e^{-\frac{(\frac{1}{s_{ij}})^2}{\tau^2}}$
  - Similarity to future publications (I_f)
  - Final score: $trendiness = \frac{I_p N_f}{I_f N_p}$

Conclusion

- It might be a good idea to try things out by using interactive visualization
- Future directions
  - Move from (multi) document level to text level approaches
  - Combine these approaches
  - Enable users to understand uncertainties visually



References

- Bosch, Harald, et al. "Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering." *Visualization and Computer Graphics, IEEE Transactions on* 19.12 (2013): 2022-2031.
- Bühler, Thomas, and Matthias Hein. "Spectral clustering based on the graph p-Laplacian." *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009.
- Chen, Chaomei. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." *Journal of the American Society for information Science and Technology* 57.3 (2006): 359-377.
- Councill, Isaac G., C. Lee Giles, and Min-Yen Kan. "ParsCit: an Open-source CRF Reference String Parsing Package." *LREC*. 2008.
- Dou, Wenwen, et al. "Leadline: Interactive visual analysis of text data through event identification and exploration." *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012.
- Fried, Daniel, and Stephen G. Kobourov. "Maps of computer science." *Pacific Visualization Symposium (PacificVis), 2014 IEEE*. IEEE, 2014.
- Hoferlin, B., et al. "Inter-active learning of ad-hoc classifiers for video visual analytics." *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*. IEEE, 2012.
- Lee, Bongshin, et al. "Understanding research trends in conferences using PaperLens." *CHI'05 extended abstracts on Human factors in computing systems*. ACM, 2005.
- Ley, Michael. "The DBLP computer science bibliography: Evolution, research issues, perspectives." *String Processing and Information Retrieval*. Springer Berlin Heidelberg, 2002.
- MacEachren, Alan M., et al. "Senseplace2: Geotwitter analytics support for situational awareness." *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011.
- Marcus, Adam, et al. "Twitinfo: aggregating and visualizing microblogs for event exploration." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2011.
- Rayson, Paul, and Roger Garside. "Comparing corpora using frequency profiling." *Proceedings of the workshop on Comparing Corpora*. Association for Computational Linguistics, 2000.
- Wei, Furu, et al. "Tiara: a visual exploratory text analytic system." *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.
- Zhang, Jian, Chaomei Chen, and Jiexun Li. "Visualizing the intellectual structure with paper-reference matrices." *Visualization and Computer Graphics, IEEE Transactions on* 15.6 (2009): 1153-1160.