Visual Analysis of Relatedness in Dynamically Changing Repositories

Coupling Visualization with Machine Processing for Gaining Insights into Massive Data

Vedran Sabol

Know-Center GmbH

gefördert durch das Kompetenzzentrenprogramm

Competence Centers for Excellent Technologies

Overview

- Motivation
  - gain understanding of large amounts of complex information
- Traditional ways: knowledge discovery and information visualisation
- Visual analytics: what it is and why it is useful
- Topical-temporal analysis of text corpora
  - Algorithms
  - Topical analysis: information landscape visualisation
  - Temporal visualisation: StreamView
  - Multiple component analytical user interface

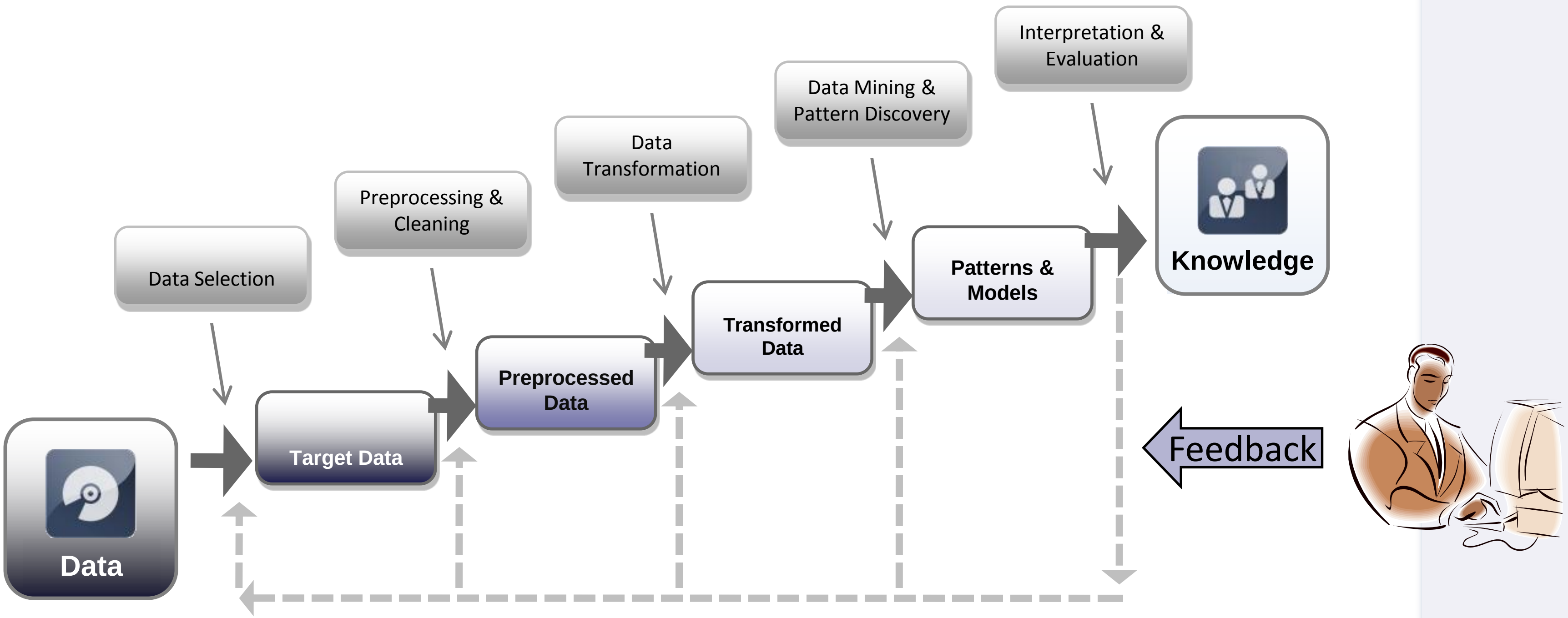
Motivation

- We are confronted with:
  - Massive amounts of information
  - Complex (multi-dimensional) knowledge objects
  - Change and the temporal dimension
  - Unstructured (i.e. text) and semi-structured data repositories
- How to utilise all this information?
  - Navigate, explore, analyse, understand
  - Unveil important facts and knowledge hidden within the data

Motivation
Analysis of Text Corpora

- Text remains an essential data type in many domains
  - Complex: unstructured, vague and ambiguous (synonyms, homonyms…)
  - Abstract concepts and relationships, interpretation by humans
  - Non-visual: no “natural” visual representation for text corpora
- Apply visual analytics on dynamic text corpora to detect
  - Dominant topics and their relationships
  - Major trends and events
  - Role of entities/metadata, such as locations, temporal references, persons or other objects of interest (such as historical buildings)
  - Correlations between trends, topics, and entities

Knowledge Discovery



- Knowledge Discovery Process [Fayyad, 1996]
  - Mainly an automatic approach consisting of a chain of processing steps
  - Goal: discovery of new, relevant, previously unknown patterns and relationships in data

Knowledge Discovery

Limitations of Automatic Analysis

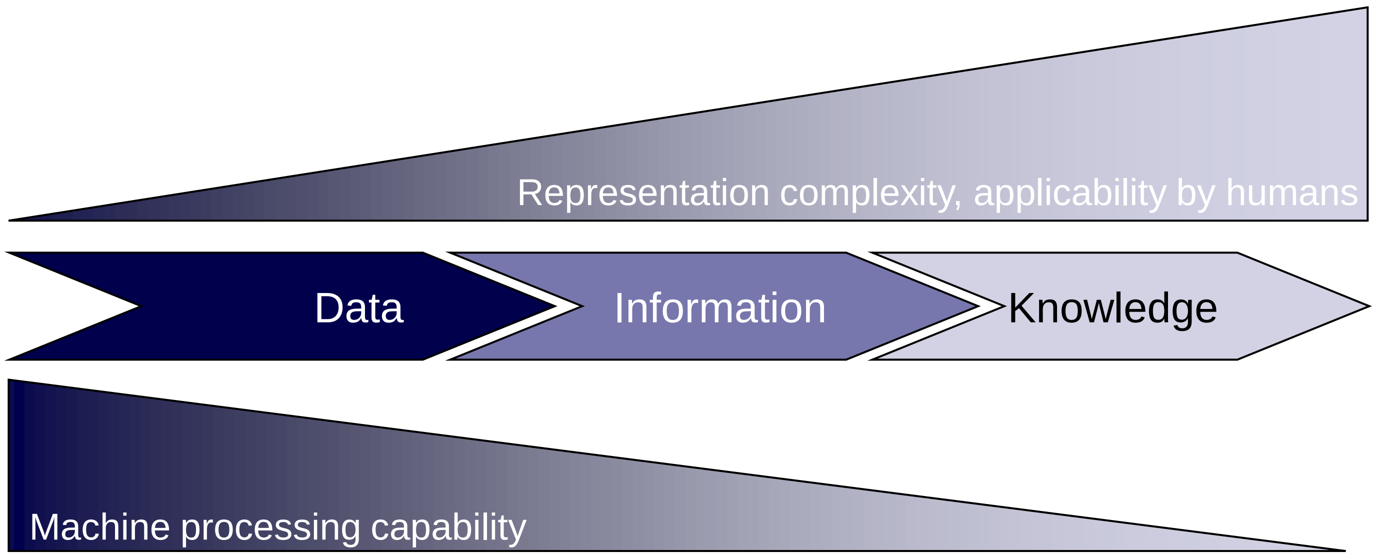
- Machines are very powerful
  - Automatic processing methods for huge data sets
  - Exponential growth of computer-performance since 60 years
    - Moor‘s Law: continues until 2020, 2030… ?
  - Distributed computing: Grid, Cloud, …
- Nevertheless, machines still behind humans in
  - **Identification of complex patterns and relationships**
  - Wide knowledge and experience
  - Abstract thinking
  - …

Visualisation

- Human visual apparatus is an extremely efficient „processing machine“
  - Enormous amounts of information are transferred by the visual nerve into the brain cortex
  - Visual cortex is unbeatable in recognition of complex patterns
  - Pre-attentive processing: no need to focus our attention
    - Processing time < 200 - 250ms, independently of the noise amount
- Visualisation definition
  - Use of human visual system, supported by computer graphics, to analyse and interpret large amounts of data
  - Graphical representation of data, information and knowledge
  - …

Visualisation

Data – Information - Knowledge



- What do we need?
  - Data/Scientific Visualisation
  - Information Visualisation
  - Knowledge Visualisation

Visualization
Data – Information - Knowledge

- Data: formal representation of raw, basic facts
  - Defined format (numbers, dates, strings) and meaning (no interpretation required)
  - „3162“: hotel room number (not a telephone number)
- Information: result of processing and interpretation of data
  - May not have a fixed format (unstructured or semi-structured)
  - Meaning determined by interpretation within some context
  - “The mouse is small and light” – a computer or a field mouse?
- Knowledge: identified, organized and as valid recognized information
  - Representation of reality through abstract, domain-dependent models
  - Complex formal conceptual systems, such as Ontologies

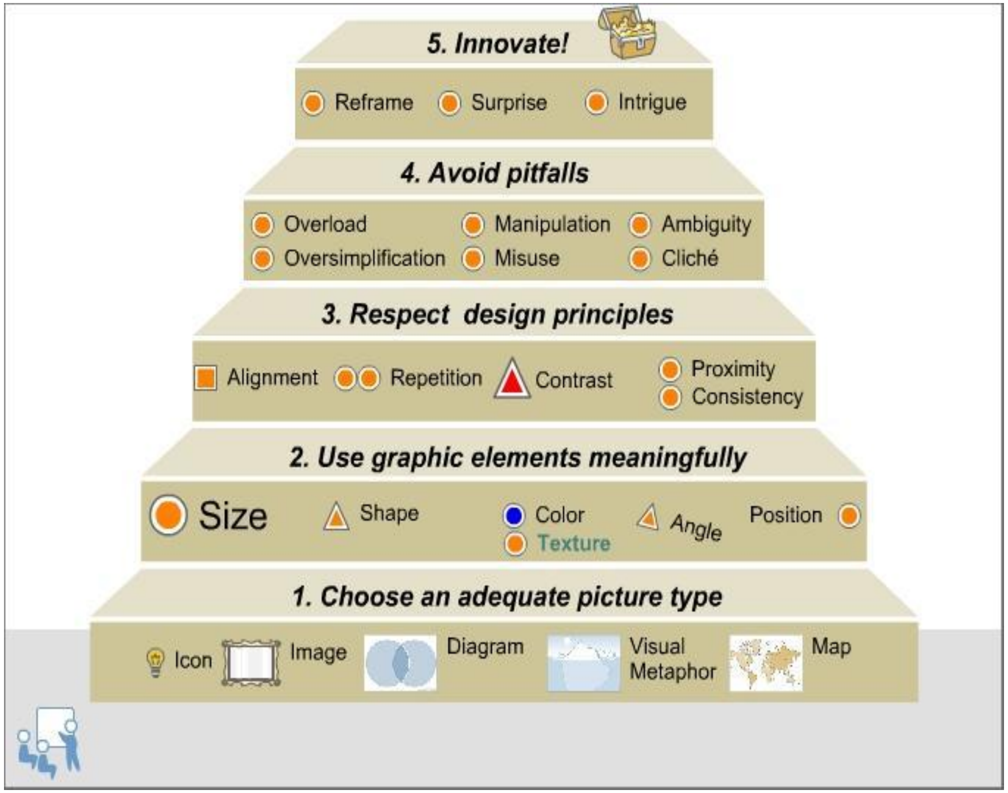
Visualization

Fundamental categories of visual representation

- Formalisms: abstract schematic representations
  - Defined by a designer, users must learn how to read and interpret
  - Example: Percentage is represented by an arc
- Metaphors: based on a real-world equivalent
  - Intuitive, user understands the meaning through building analogies
  - Example: geographic map metaphor represents similarity in non-spatial data
- Models: based on mental representations of real physical world
  - Data typically has a natural representation in the real world
  - Examples: Virtual 3D worlds, visualization of sensory data in 3D

Knowledge Visualization

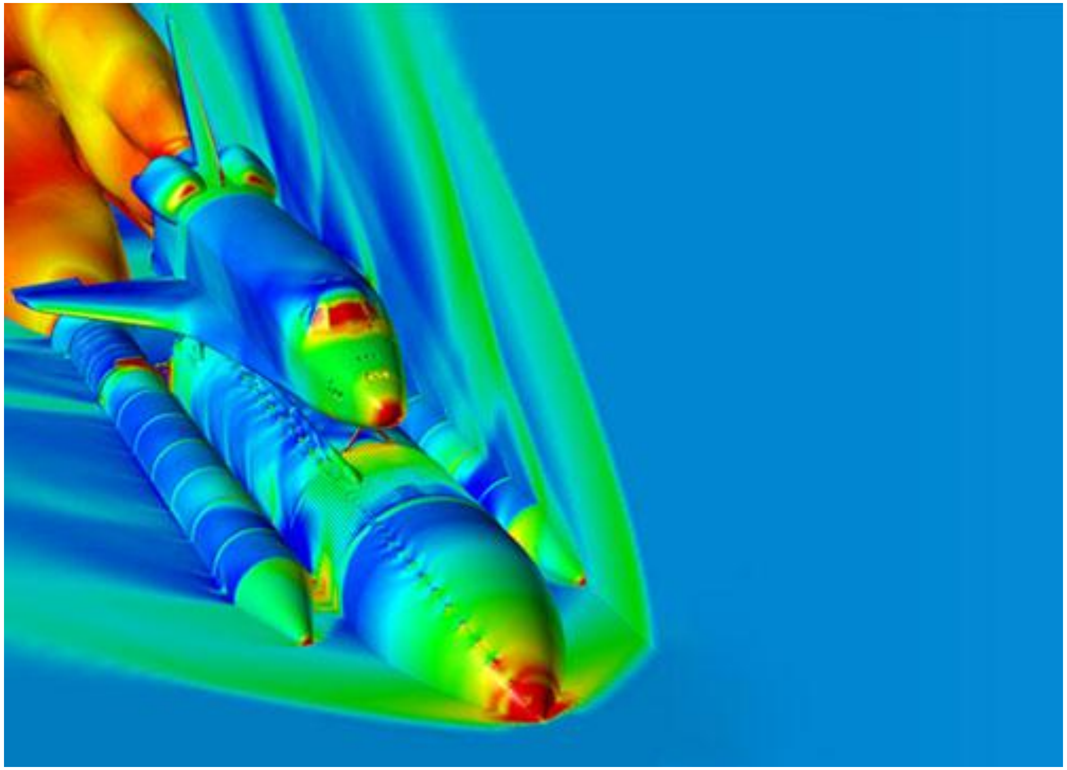
- Knowledge Visualization is about using visual representations to express and transfer existing knowledge between people [Eppler]
  - Use of metaphors and formalisms is common (static and interactive forms)



Stairs of Visualisation [Eppler]
(let's focus KV software)

Scientific/Data Visualization

- Visualization of raw data (simulation or sensory data)
  - (usually) have a natural representation in the real, physical world
- Applications in physics, medicine, astronomy, industry, …



Pressure coefficients [NASA]

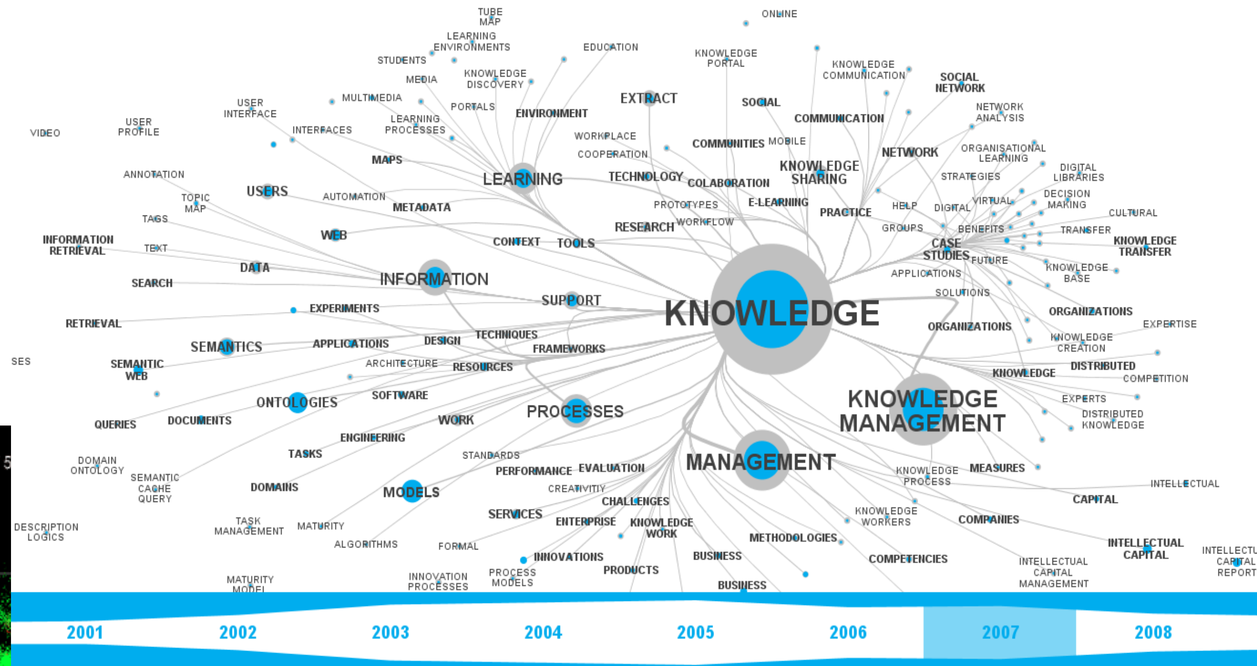
Information Visualization

- Interactive visualization of abstract information
  - No “natural” representation: use metaphors and formalisms
- Goal: identifying patterns and relationships
  - Explorative analysis
  - InfoVis Mantra: „overview first - zoom and filter - details on demand“ [B. Shneiderman]
- IV applicable on:
  - Content: Text and Multi-Media
  - Multidimensional data sets
  - Structures: Hierarchies, Networks and Graphs
  - Temporal Information
  - Geo-spatial Information
  - …

Information Visualization

Examples

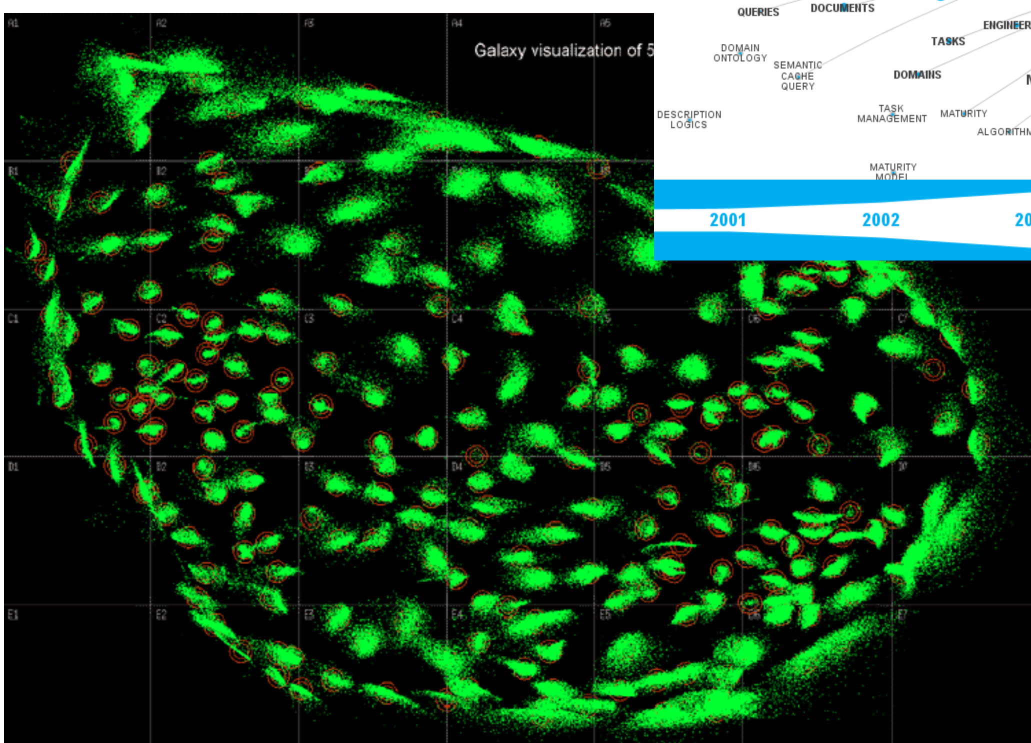
www.know-center.at



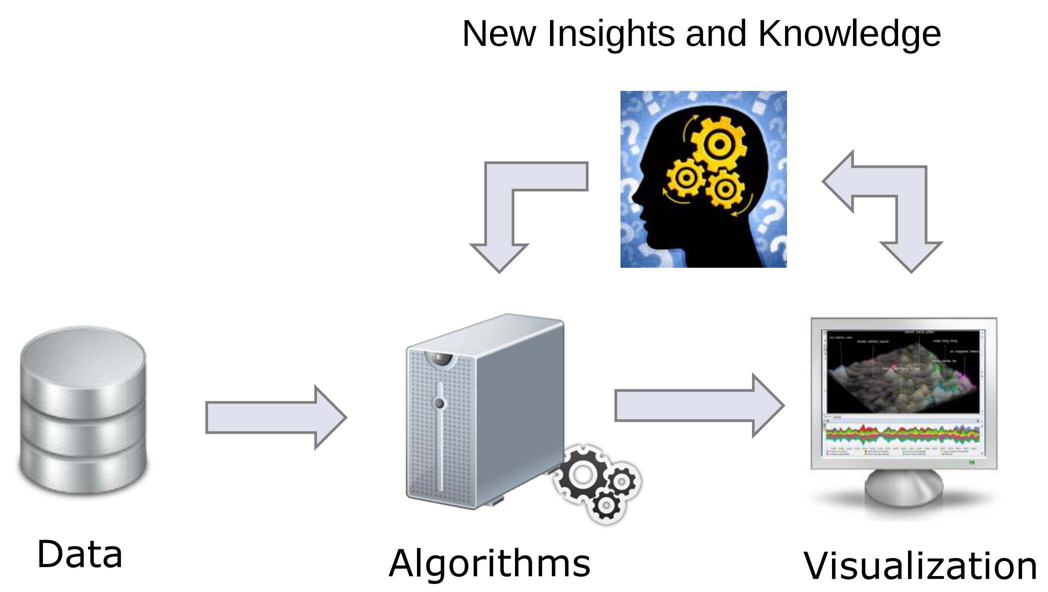
Concept Networks [Kienreich] (Know-Center)

IN-SPIRE Galaxies

[Wise] (PNNL)



Visual Analytics

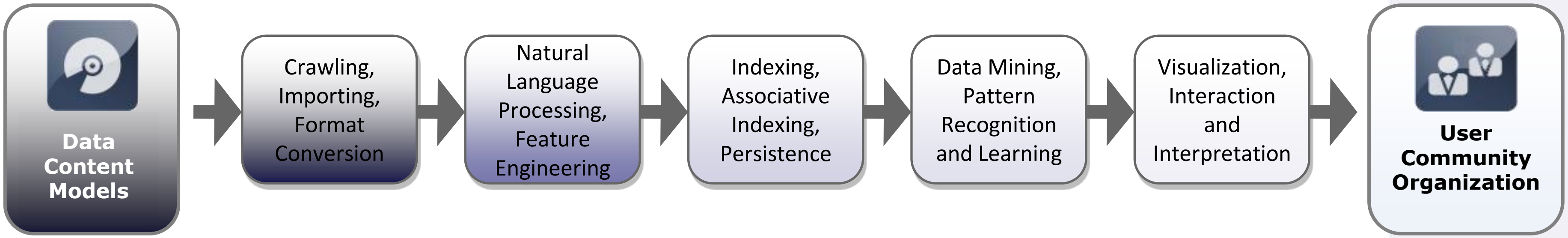


- Combines automatic methods with interactive information(/data) visualisation to get the best of both worlds [Keim 2008, Thomas 2005]
- Goal: support analytical reasoning and deriving of new knowledge from data
  - Integrates humans in the analytical process
  - Provides means for explorative analysis

Visual Analytics

- Main Idea (Mantra): *"analyse first – show the important – zoom, filter and analyse further – details on demand"* [Keim 2008]
  - Initial analysis and visual pattern recognition
  - Posing a hypothesis
  - Further analysis steps (automatic and interactive)
  - Confirmation or rejection of the hypothesis: new facts
    - Confirm the expected, discover the unexpected
- Challenges [Keim 2009]
  - Balance between automatic and interactive analysis
  - Design of effective VA workflows
  - Data quality
  - Scalability

Text Analysis Pipeline



- Resembles the Knowledge Discovery process closely
- Includes visual and automatic methods
- Convert text information into a visual representation to identify
  - Dominant topics and their relationships
  - Major trends and events
  - Role of entities/metadata, such as locations, persons, organisations…
  - Correlations between trends, topics, and entities

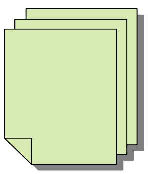
Klieber, W., Sabol, V., Muhr, M., Kern, R., Öttl, G., Granitzer, M. (2009) Knowledge Discovery Using the KnowMiner Framework, IADIS International Conference Information Systems

Pattern Recognition in Text Data

Projection Algorithms

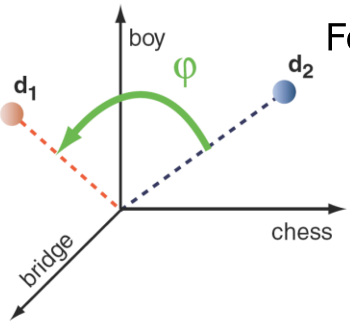
- Text is described by a very large amount of features
- How to visualize large, high-dimensional data sets?
- Projection into a „smaller“ (2D) visualization space which can be understood by users
  - Navigation and explorative analysis in the projection space
- Dimensionality reduction (ordination) techniques
  - Projection of the high-dimensional space into a lower dimensional one
  - Preservation of distances/similarities
  - Usability and aesthetics play an important role

Projection Algorithm



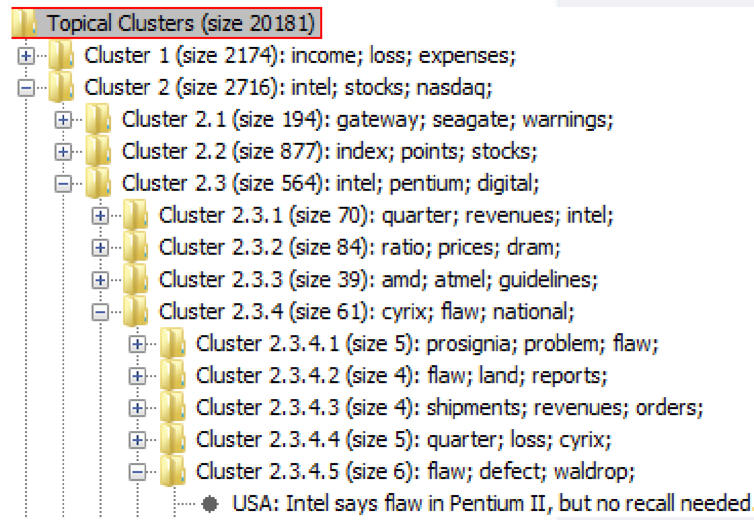
Data (Text)

Feature Engineering



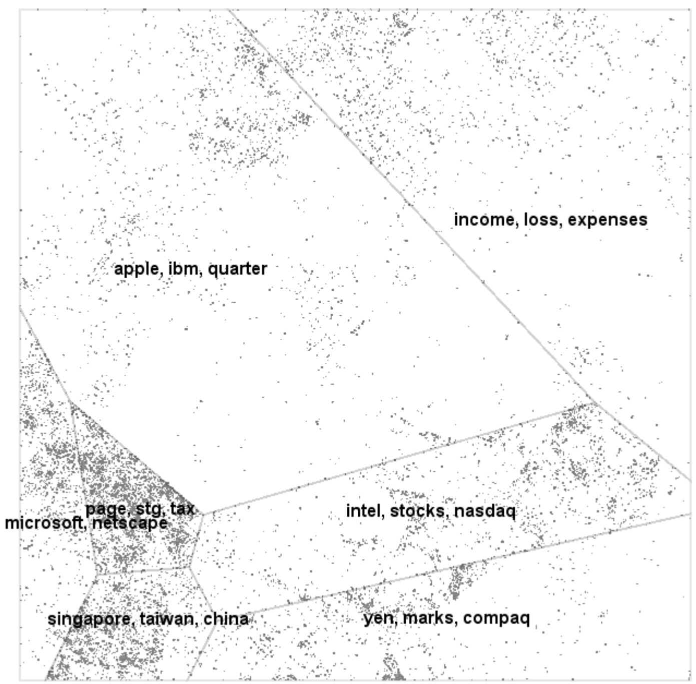
Feature Vectors

Clustering


Topical Clusters (size 20181)
Cluster 1 (size 2174): income; loss; expenses;
Cluster 2 (size 2716): intel; stocks; nasdaq;
Cluster 2.1 (size 194): gateway; seagate; warnings;
Cluster 2.2 (size 877): index; points; stocks;
Cluster 2.3 (size 564): intel; pentium; digital;
Cluster 2.3.1 (size 70): quarter; revenues; intel;
Cluster 2.3.2 (size 84): ratio; prices; dram;
Cluster 2.3.3 (size 39): amd; atmel; guidelines;
Cluster 2.3.4 (size 61): cyrix; flaw; national;
Cluster 2.3.4.1 (size 5): prosignia; problem; flaw;
Cluster 2.3.4.2 (size 4): flaw; land; reports;
Cluster 2.3.4.3 (size 4): shipments; revenues; orders;
Cluster 2.3.4.4 (size 5): quarter; loss; cyrix;
Cluster 2.3.4.5 (size 6): flaw; defect; waldrop;
USA: Intel says flaw in Pentium II, but no recall needed

Hierarchy: “virtual table of contents”

Layouting



Hierarchical geometry: spatial proximity conveys relatedness

Projection Algorithm

- Input: term vectors, base area (rectangle)
- Output: hierarchy of nested areas, 2D document positions
- Recursive Algorithm
  - Aggregation: k-means clustering, labelling using highest weight features
  - Similarity layout: force-directed placement, inscribing into area
  - Area subdivision: Voronoi diagrams
  - For each cluster: cluster size > threshold?
    - Yes: apply algorithm on cluster
    - No: layout documents

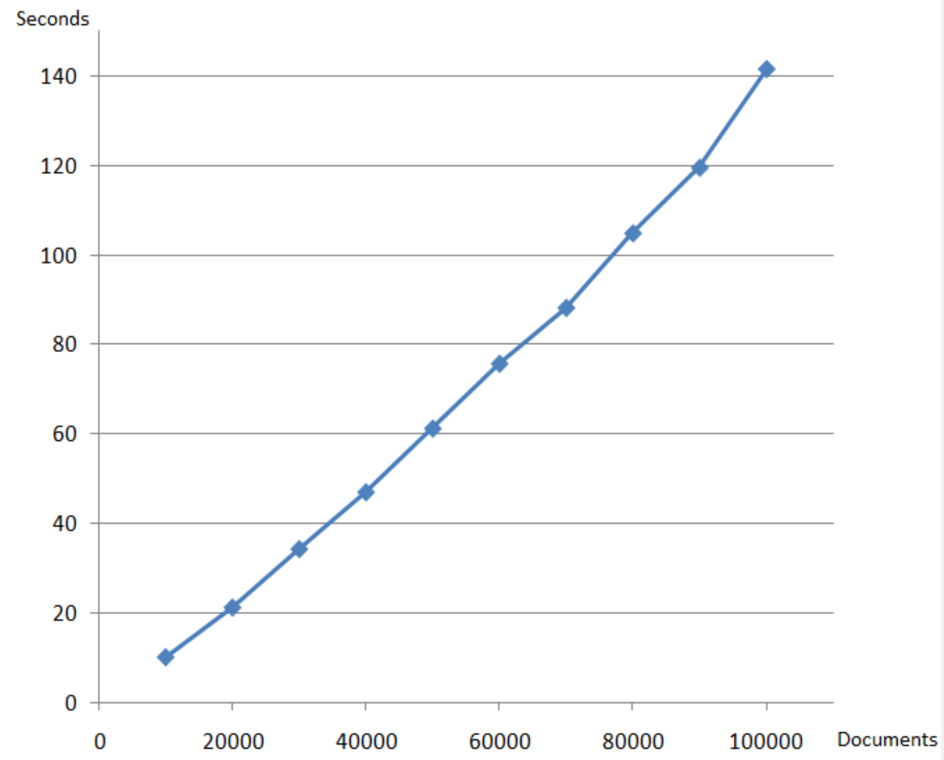
Andrews, K., Kienreich, W., Sabol, V., Becker, J., Kappe, F., Droschl, G., Granitzer, M., Auer, P., Tochtermann, K., The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities, Palgrave Journals: Information Visulization, 2002.

Sabol, V., Syed, K.A.A., Scharl, A., Muhr, M., Hubmann-Haidvogel, A., Incremental Computation of Information Landscapes for Dynamic Web Interfaces, Proceedings of the 10th Brazilian Symposium on Human Factors in Computer Systems, 2010.

Muhr, M., Sabol. V., Granitzer, M., Scalable Recursive Top-Down Hierarchical Clustering Approach with implicit Model Selection for Textual Data Sets, IEEE 7th International Workshop on Text-based Information Retrieval in Proceedings of DEXA'10 2010.

Projection Algorithm
Properties

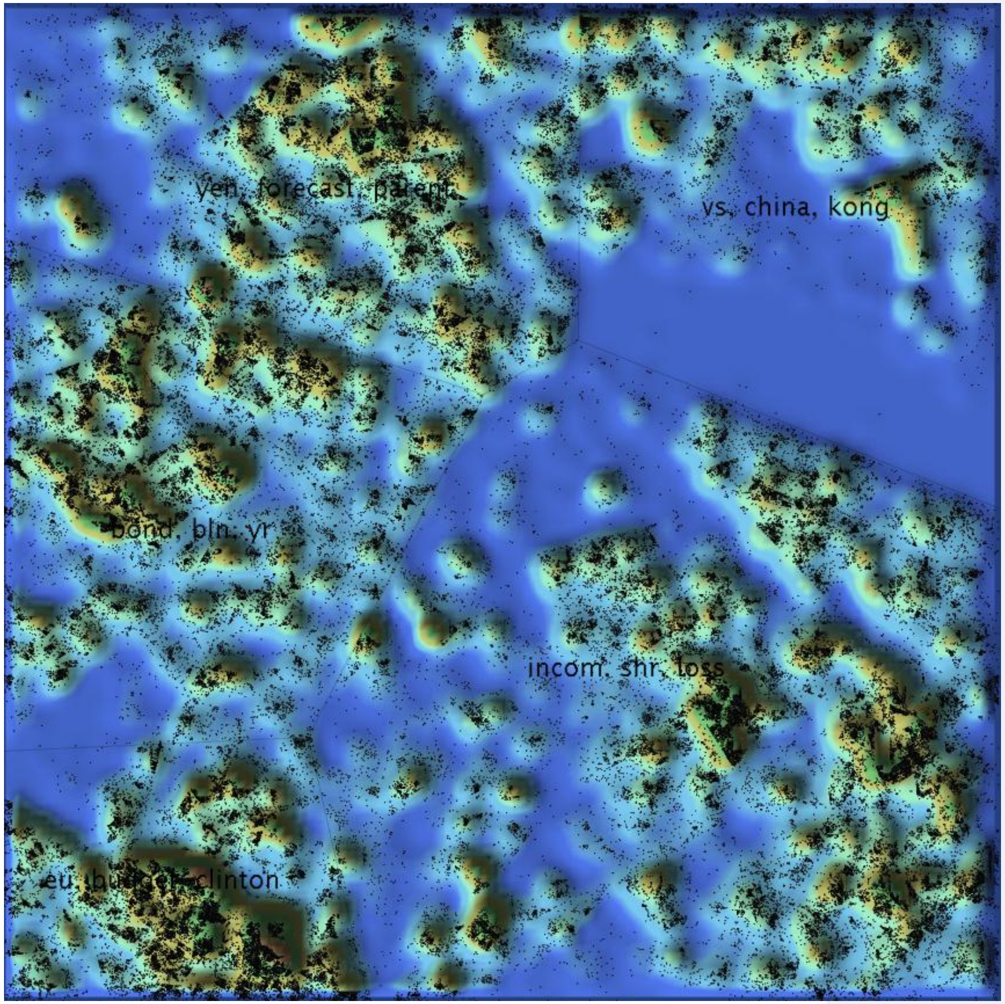
- Scalability
  - Time and space complexity: O(n*log(n))
- Hierarchical geometry
  - Navigation and exploration
- Incremental
  - Incorporate data changes into an existing layout
- Parameterisable
  - Adaptable to different data types
  - Can be tuned to produces layouts suitable for visualisation



Information Landscape

Relatedness (Topical Similarity)

- Proximity expresses relatedness
- Hills represent groups of similar data elements
  - Height indicates size
  - Compactness indicates topical cohesion
- Labels capture essence of undelaying data
  - Orientation and navigation

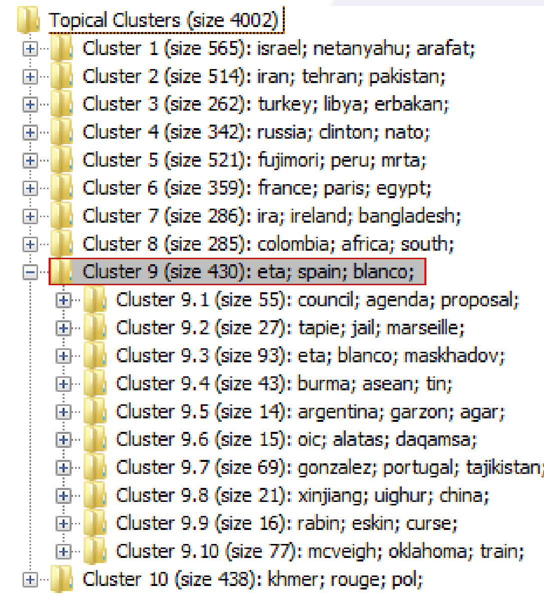


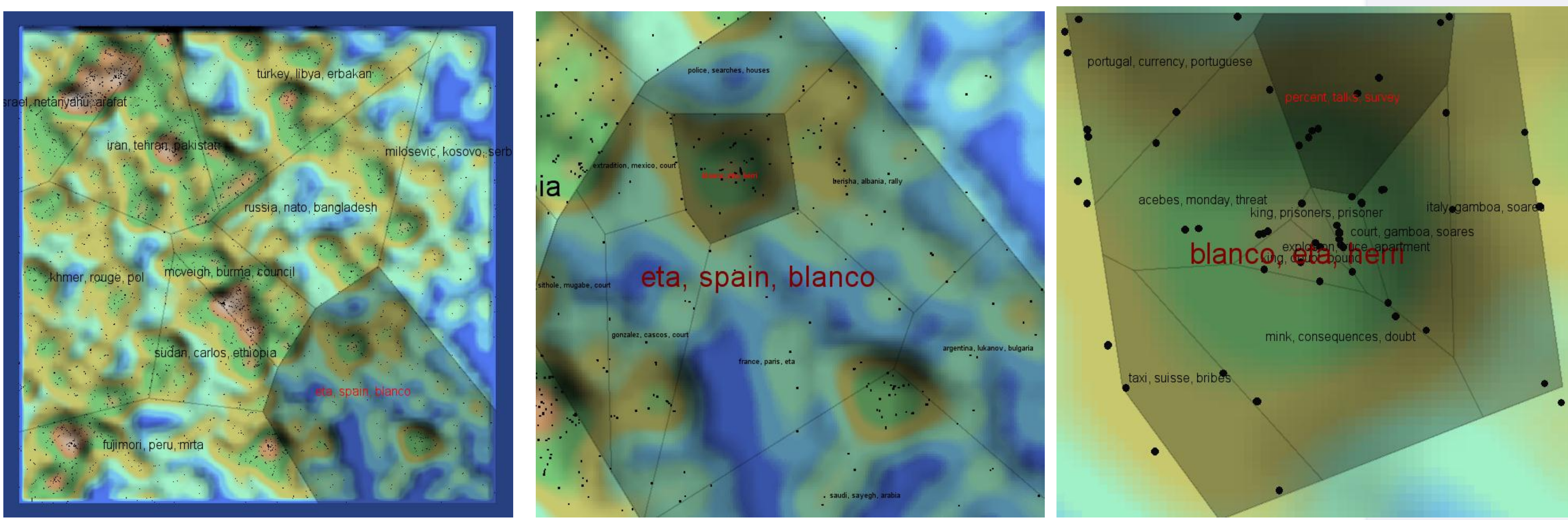
400.000 documents (from RCV1)

Hierarchical Information Landscape

Navigation and Orientation

- Conveys relatedness and hierarchy
- Level of detail-sensitive navigation and orientation
  - Animated transitions: auto-focus on the chosen cluster



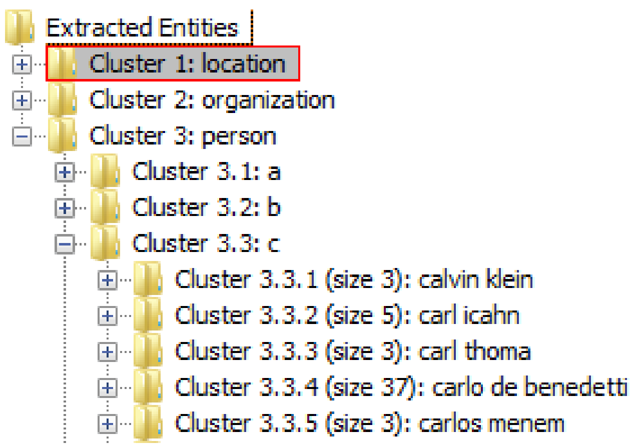


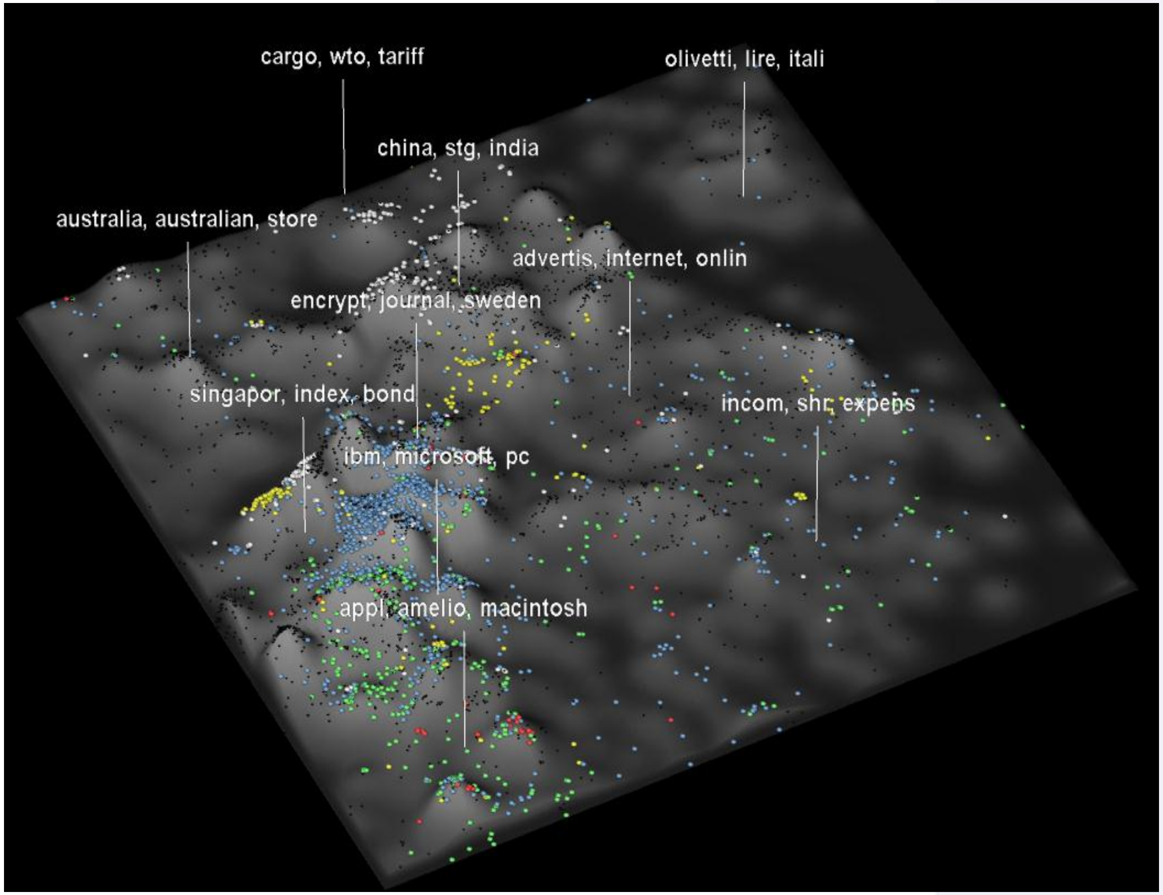
Granitzer, M., Kienreich, W., Sabol, V., Andrews, K., Klieber, W., Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories, InfoVis '04, the tenth annual IEEE Symposium on Information Visualization, 2004.

Information Landscape

Faceted Metadata

- Correlations between topics and entities/metadata



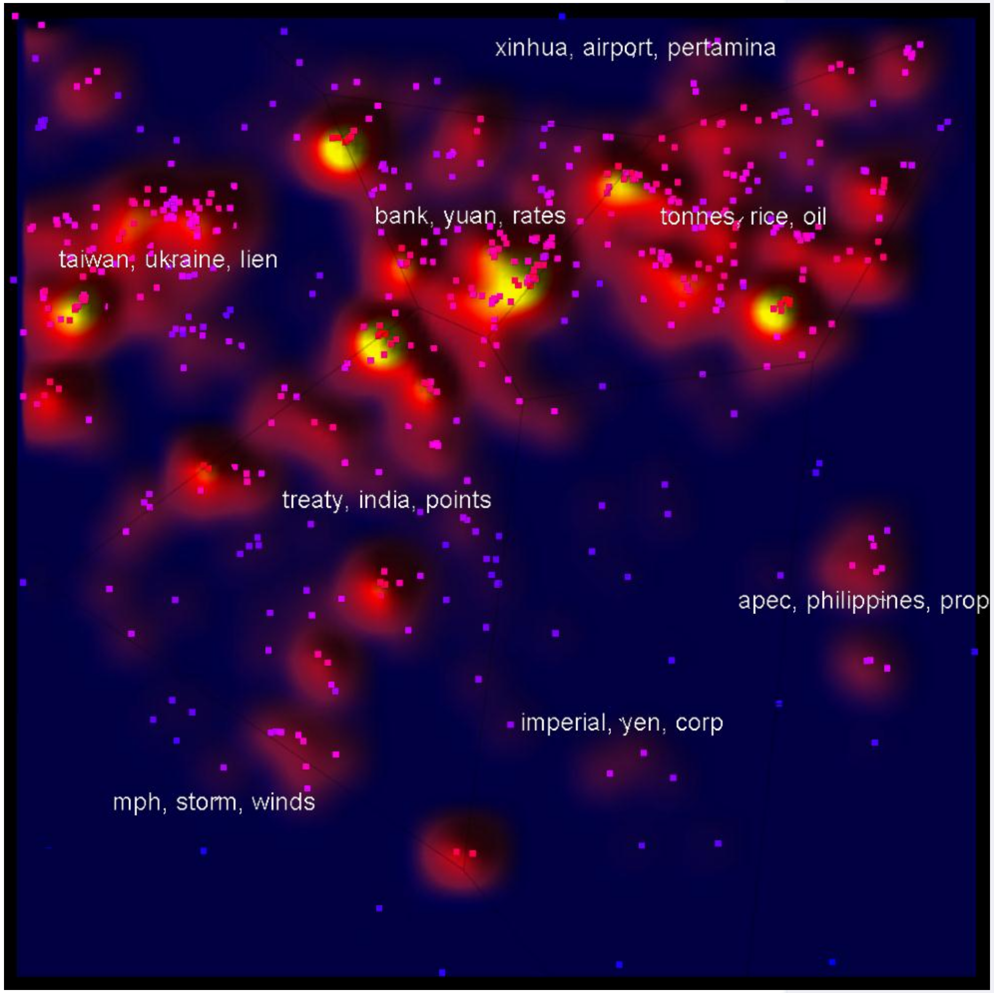


- Facetted metadata clusters
  - Colour-coded

Information Landscape

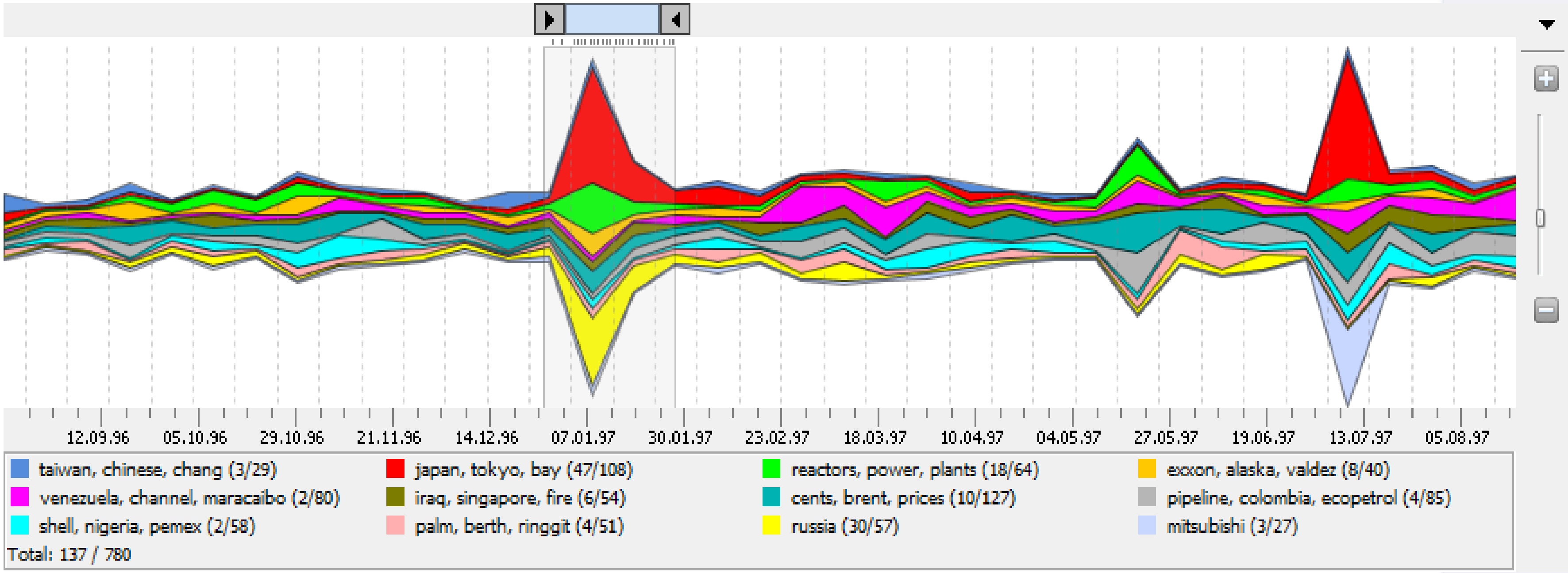
Numerical Information

- Items: map onto transparency, colour, size
- Map: topography, heat map
- Applicable to uncertainty
  - Height and colour encode certainty (information quality)



StreamView

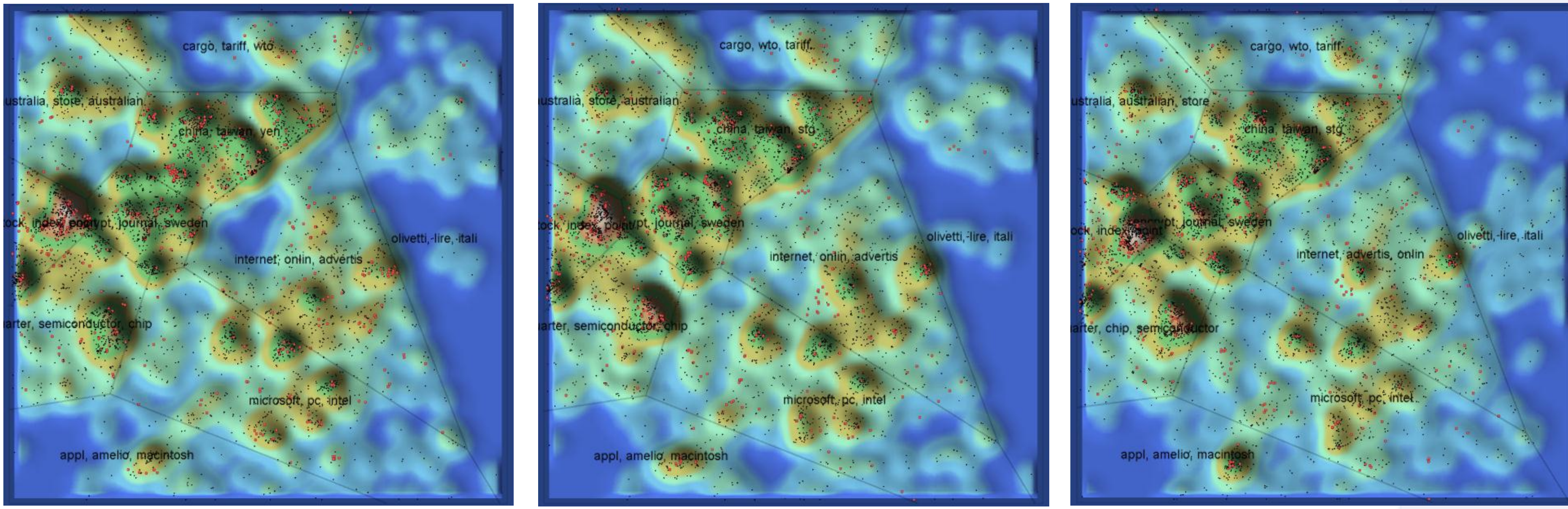
Temporal Information



- Detect
  - Trends and changes in topical and faceted metadata clusters
  - Temporal correlations between clusters

Dynamic Information Landscape

Incremental Integration of Changes



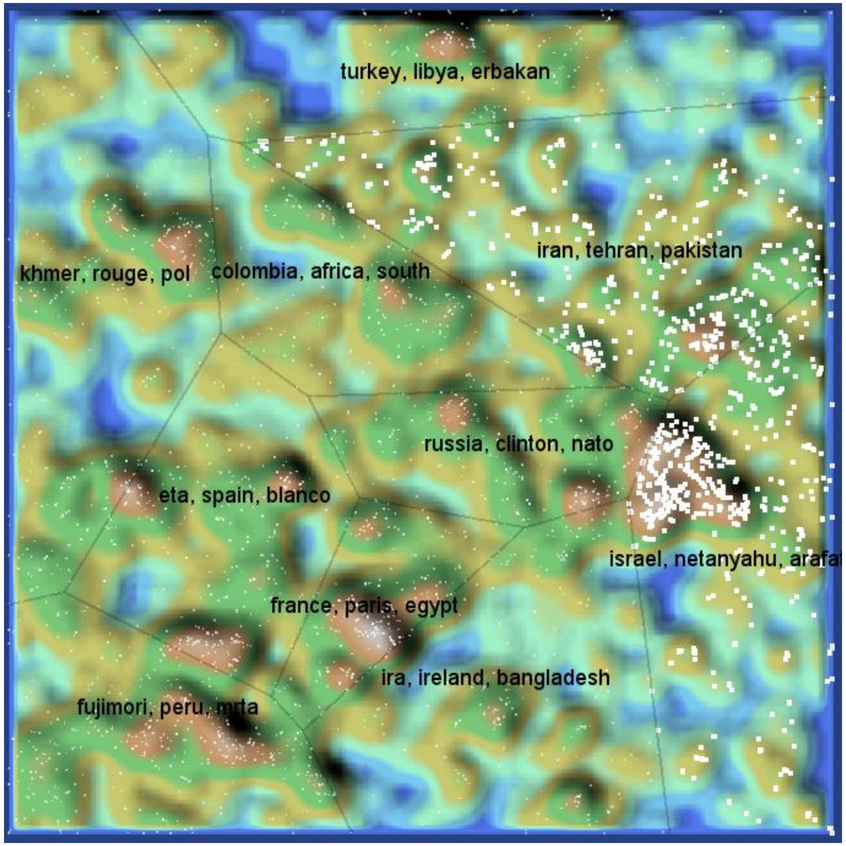
- Change in the layout corresponds to change in the data
  - User retains recognition and orientation through unchanged parts of the topography

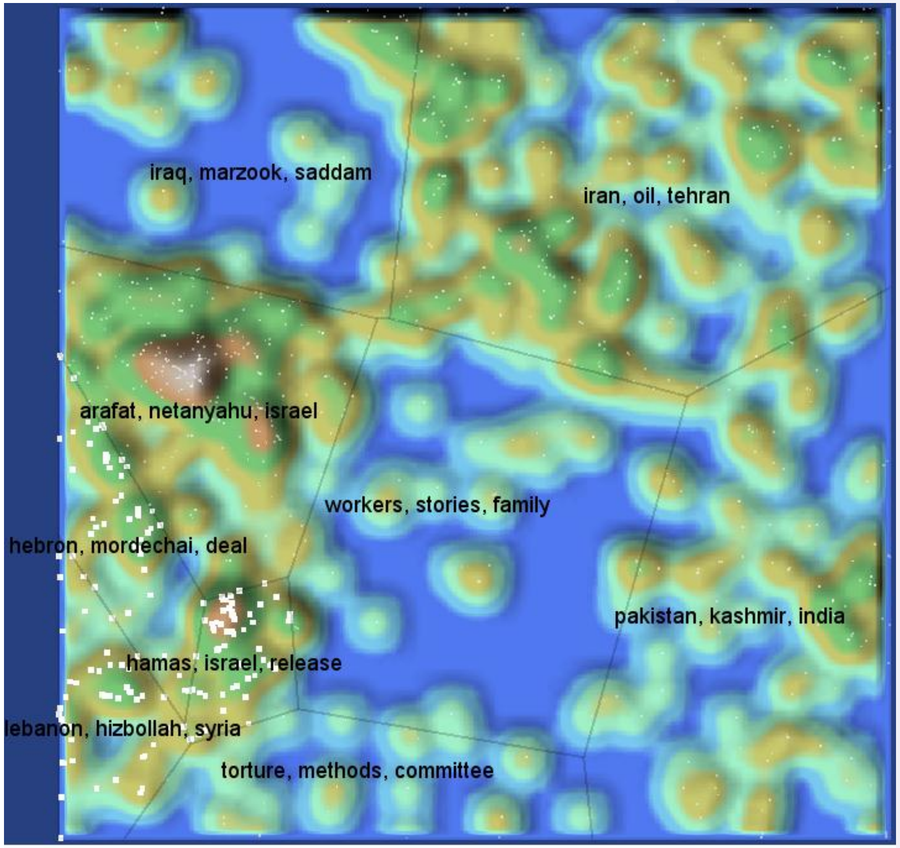
Sabol, V., Scharl, A. (2008) Visualizing Temporal-Semantic Relations in Dynamic Information Landscapes, GeoVisualization of Dynamics, Movement and Change Workshop at the AGILE 2008 Conference, Spain.

Sabol, V., Kienreich, W. (2009) Visualizing Temporal Changes in Information Landscapes, EuroVis 2009.

Visual Scatter/Gather

Drill Down





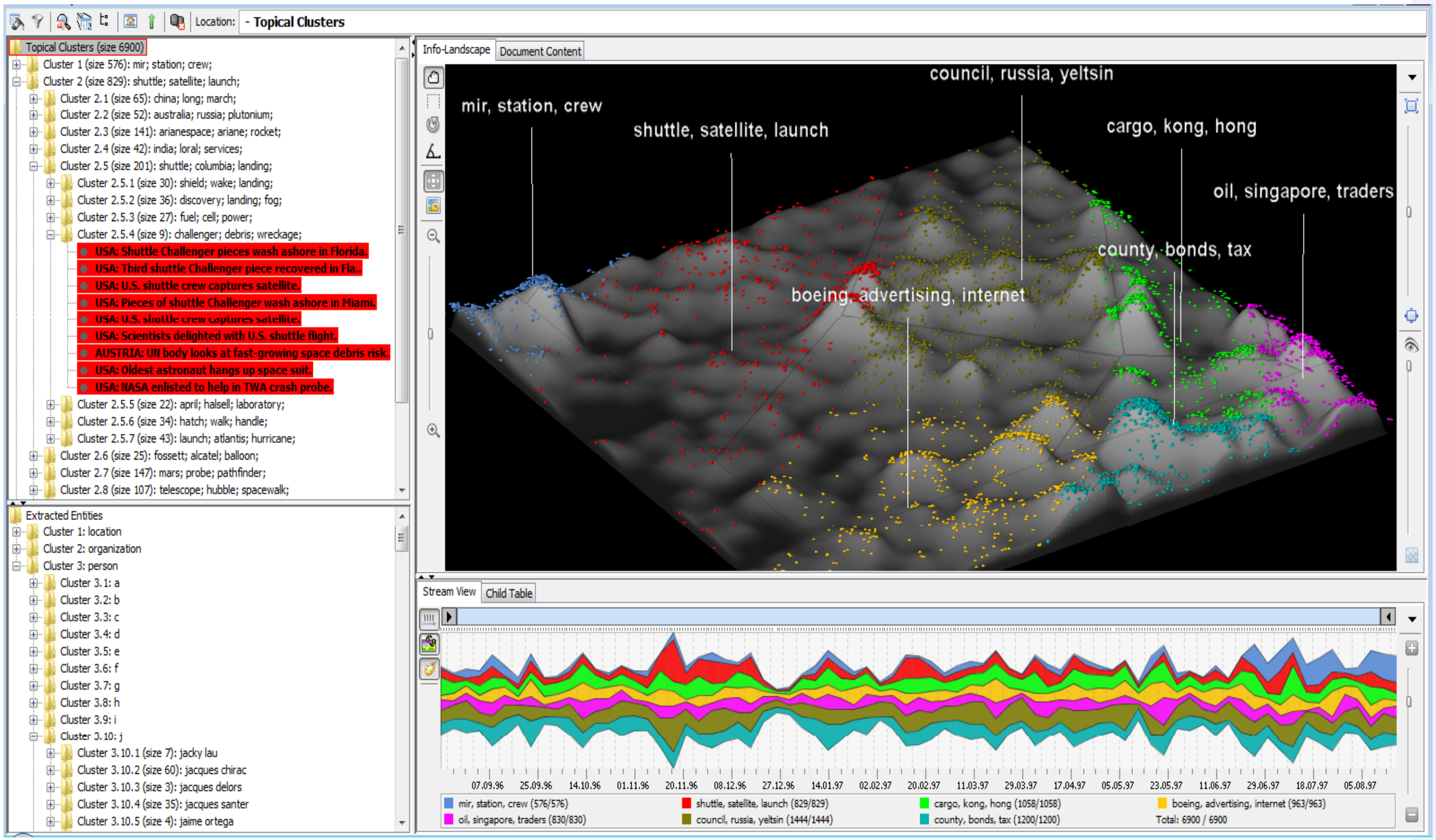
- Identify and select relevant parts of the data set
- Retrigger analysis to focus on the chosen subset

Multiple Coordinated Views
Multiple Data Aspects

- Multiple visualizations “fused” into a single, coherent user interface
- Each visualization is designed to presents a different data aspect
  - Relatedness, time information, hierarchical structure …
- MCVs enable simultaneous analysis over multiple data aspects
- Coordination: interactions in one component influence all others
  - Colours and transparency
  - Icons
  - Size
  - Selection
  - Navigation
  - Visibility
  - …

Topical-Temporal-Metadata Analysis

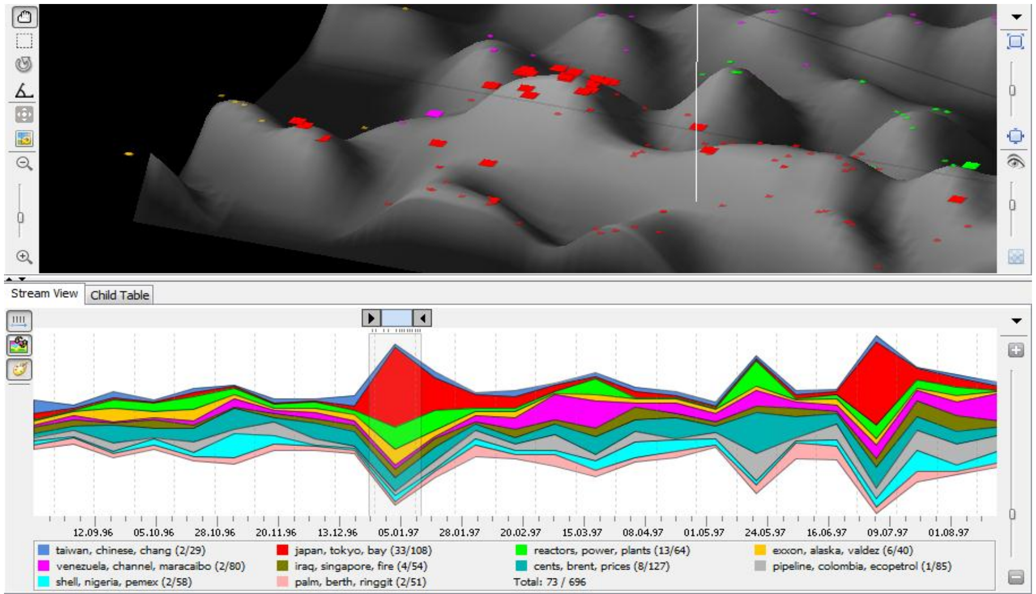
Multiple Coordinated View Interface

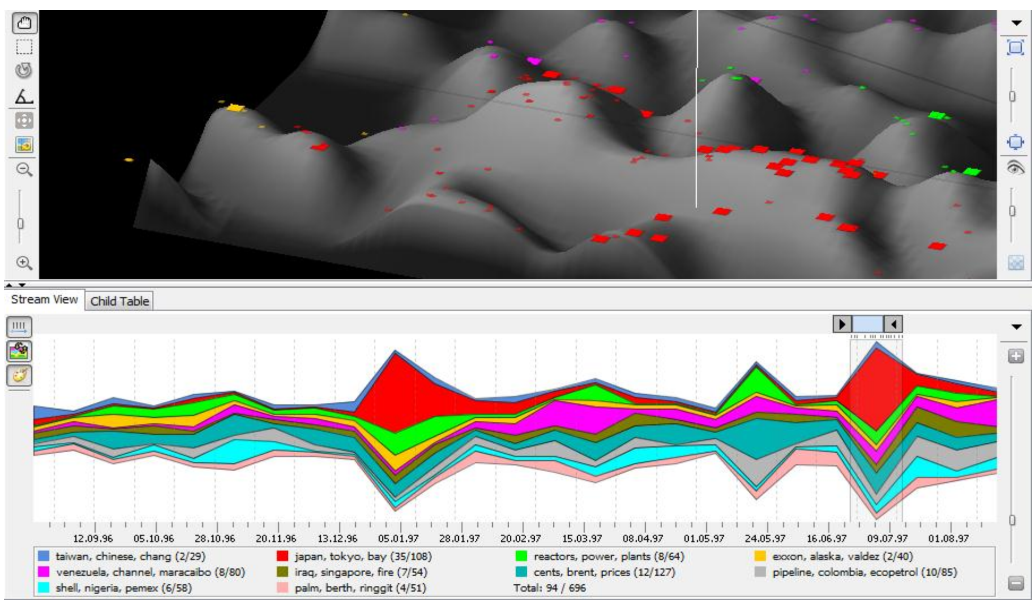


Sabol, V., Granitzer, M., Kienreich, W. (2007) Fused Exploration of Temporal Developments and Topical Relationships in Heterogeneous Data Sets, in Proceedings of the 11th International Conference Information Visualisation (IV'07), IEEE.

Sabol, V., Kienreich, W., Muhr, M., Klieber, W., Granitzer, M. (2009) Visual Knowledge Discovery in Dynamic Enterprise Text Repositories, Proceedings of the 13th International Conference on Information Visualisation (IV09), IEEE Computer Society.

Topical-Temporal Analysis - Example





“Japan, Tokyo, Bay” cluster (red)

- 2 temporal peaks
- Topically separated (different hills)

Hypothesis: two different events

Analysis for validation:

- Inspection
- Searching + highlighting
- Correlating metadata

www.know-center.at

Usability Evaluation

- Formal Experiments with 10-15 users
  - Measure user performance
  - Discover usability issues
  - Optimise the interface
- Multiple Coordinated Views
  - Hierarchical Landscape + Tree View vs. Landscape only: better task completion rates with MCV
  - Temporal-topical analysis using Landscape + StreamView vs. Table + StreamView: two visualisations performed better
- Navigation: automatic vs. manual zooming/panning – mixed results
- Symbol design: Complex symbols bad, colour coding recognisable

KNOW Center

www.know-center.at

Applications

- Client-server system developed (with an industry partner)
  - Integrates developed algorithmic and visual methods
  - Applied on text repositories with over 10^6 documents
  - Feedback from real users
- Applications
  - Explorative analysis of governmental document repositories
  - Patent analysis in the industry
    - Long term monitoring of technology fields
  - Interest from industry (media and technology)

Conclusion

- Visual analytics combines automatic processing and interactive visualisation providing advantages of both
  - Tightly integrates humans in the analytical process
- Demonstrated an approach for topical-temporal-metadata analysis of large text corpora
  - Usability experiments and applications in the industry confirm the viability of the approach
- Challenges
  - Text mining methods are domain and task sensitive
    - Extensive testing and tuning necessary
  - Visual methods target experts (not “consumers”)
    - Obtaining meaningful results not straightforward

Thank You!

Questions?

Dr. Vedran Sabol
Lead Visualisation Group
Know-Center GmbH
Inffeldgasse 13
8010 Graz

+43 316 873 30850
vsabol@know-center.at
www.know-center.at