Chapter 7
Visual Analysis and Knowledge Discovery for Text

Christin Seifert, Vedran Sabol, Wolfgang Kienreich, Elisabeth Lex, and Michael Granitzer

Abstract Providing means for effectively accessing and exploring large textual data sets is a problem attracting the attention of text mining and information visualization experts alike. The rapid growth of the data volume and heterogeneity, as well as the richness of metadata and the dynamic nature of text repositories, add to the complexity of the task. This chapter provides an overview of data visualization methods for gaining insight into large, heterogeneous, dynamic textual data sets. We argue that visual analysis, in combination with automatic knowledge discovery methods, provides several advantages. Besides introducing human knowledge and visual pattern recognition into the analytical process, it provides the possibility to improve the performance of automatic methods through user feedback.

7.1 Introduction

The already huge amount of electronically available information is growing further at an astonishing rate: an IDC study [12] estimates that by 2006 the amount of digital information exceeded 161 Exabyte, while an updated forecast [13] estimates that by 2012 the amount of information will double every 18 months. While retrieval tools excel at finding a single, or a few relevant pieces of information, scalable analysis techniques, considering large data sets in their entirety, are required when a holistic view is needed.

Knowledge discovery (KD) is the process of automatically processing large amounts of data to identify patterns and extract useful new knowledge [9].

C. Seifert (✉) • M. Granitzer
University of Passau, 94030 Passau, Germany
e-mail: christin.seifert@uni-passau.de; michael.granitzer@uni-passau.de

V. Sabol • W. Kienreich • E. Lex
Know-Center Graz, Inffeldgasse 13/6, A-8010 Graz, Austria
e-mail: vsabol@know-center.at; wkien@know-center.at; elex@know-center.at

A. Gkoulalas-Divanis and A. Labbi (eds.), *Large-Scale Data Analytics*,
DOI 10.1007/978-1-4614-9242-9_7, © Springer Science+Business Media New York 2014

KD was traditionally applied on structured information in databases; however, as information is increasingly present in unstructured or weakly structured form, such as text, adequate techniques were developed. The shift from large, static, homogeneous data sets to huge, dynamic, heterogeneous repositories necessitates approaches involving both automatic processing and human intervention. Automatic methods put the burden on machines, but despite algorithmic advancements and hardware speed-ups, for certain tasks, such as pattern recognition, human capabilities remain unchallenged.

Information visualization techniques rely on the powerful human visual system, which can recognize patterns, identify correlations and understand complex relationships at once, even in large amounts of data. Visualization is an effective enabler for exploratory analysis [52], making it a powerful tool for gaining insight into unexplored data sets.

Visual Analytics is an interdisciplinary field based on information visualization, knowledge discovery and cognitive and perceptual sciences, which deals with designing and applying interactive visual user interfaces to facilitate analytical reasoning [50]. It strives for tight integration between computers, which perform automatic analysis, and humans, which steer the process through interaction and feedback. Combining the advantages of visual methods with automatic processing provides effective means for revealing patterns and trends, and unveiling hidden knowledge present in complex data [23,48]. Analytical reasoning is supported based on the discovered patterns, where users can pose and test a hypothesis, provide assessments, derive conclusions and communicate the newly acquired knowledge.

Especially for large text repositories, Visual Analytics is a promising approach. Turning textual information into visual representations allows to access large document repositories using the human pattern recognition abilities. Providing Visual Analytics environments for text requires text mining and text analysis algorithms in order to extract information and metadata. Further, appropriate representations have to be devised in order to visualize the aspects interesting to the task at hand.

In this chapter, we provide an overview on visual analysis techniques for textual data sets, outline underlying processing elements and possible application scenarios. First, the general processing pipeline for Visual Analytics in text repositories is outlined in Sect. 7.2, followed by a detailed description of all the necessary steps. Second, Sect. 7.3 describes how visual representations can be used on the extracted information. Different visualizations are represented depending on the data aspect to be visualized. Section 7.3.1 describes topical overviews, Sect. 7.3.2 representations for multi-dimensional data, Sect. 7.3.3 spatio-temporal visualizations, and Sect. 7.3.4 visualization of arbitrary relations. The concept of user feedback integration, as well as examples, are covered in Sect. 7.3.5. The concept of combining multiple visualizations into one interface, as multiple coordinated views, is explained in Sect. 7.3.6. Third, Sect. 7.4 describes three applications: media analysis (Sect. 7.4.1), visual access to encyclopedias (Sect. 7.4.2) and patent analysis (Sect. 7.4.3). Finally, Sect. 7.5 concludes the chapter and provides an outlook of future developments in the field of Visual Analytics focusing on text data.

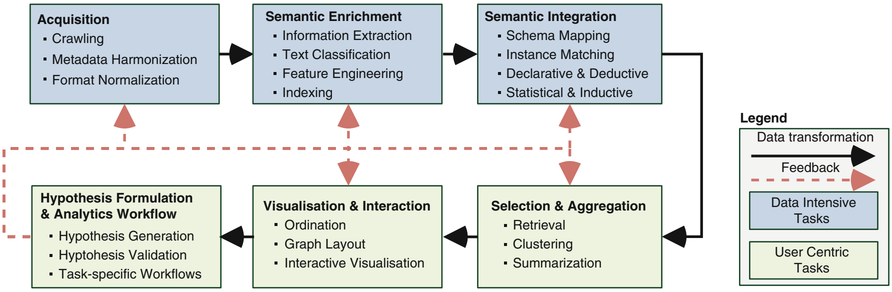

Fig. 7.1 The processing pipeline for visual analysis of text combines data-intensive tasks (*top*) and user-centric tasks (*bottom*). *Solid black lines* indicate data flows while *dashed red lines* indicate user feedback to adapt automatic processes

7.2 Processing Pipeline for Visual Analysis of Text

Visual analytics combines information visualization techniques with knowledge discovery methods in an iterative fashion. Starting from a given data set, mining techniques identify interesting, non-trivial patterns, which may provide insights on the data set. In a discovery task, where the aim is to identify new, potentially useful insights, a priori assumptions underlying the mining techniques may not be fulfilled. By visualizing the extracted patterns, humans are empowered to incorporate their background knowledge into the automatic processes through identifying wrong assumptions and erroneously identified patterns. Information visualization serves as the communication channel between the user and the mining algorithm, allowing domain experts to control the data-mining process, to rule-out wrong mining results or to focus on particularly interesting sub-samples of the data set.

Providing Visual Analytics environments for text requires text mining and text analysis algorithms, in order to extract meaningful patterns subject to visualization. The data set usually consists of a set of documents and additional metadata. Acquiring and processing these metadata is usually composed of consecutive steps resembling the traditional knowledge discovery chain. Visual analytics aims to provide users with intelligent interfaces controlling parts of these steps in order to gain new insights. Understanding individual steps is necessary to derive suitable visualizations and interactions for each step. Hence, we outline important details on the process in the following and afterwards derive potential visualizations and interactions for conducting visual analytics tasks in large text repositories. Figure 7.1 depicts an overview of the outlined process.

7.2.1 Acquisition

Acquisition includes crawling and accessing repositories to collect information, and document pre-processing, such as harmonization of metadata and conversion into a unified format. An often underestimated effort within the acquisition step is data cleaning and metadata harmonization. Data cleaning involves removing documents in awkward formats (e.g., encrypted PDFs) and data that should be omitted in further processing (e.g., binary content). Similarly, metadata harmonization ensures the correctness of data from various sources and the availability of necessary information for later semantic integration [2].

7.2.2 Semantic Enrichment

Semantic enrichment extracts domain-specific semantics from single documents and enriches each document with external knowledge. Usually, the process starts with annotating the document content with linguistic properties like part-of-speech or punctuation, or external knowledge like thesauri concepts. Annotated text then serves as basis for either extracting explicit metadata on document level, e.g., the author of a document, or to generate features representing the document in subsequent analysis steps. Annotations, metadata and features may serve as input for creating index structures to enable fast, content-based access to documents. Figure 7.2 provides an overview on these data transformations happening during semantic enrichment. In the following, we outline the most important techniques in detail.

7.2.2.1 Information Extraction

Information Extraction (IE) deals with extracting structured information from unstructured, or weakly structured, text documents using natural language processing methods [20]. IE decomposes text into building blocks, generates annotations and extracts metadata, typically employing the following methods:

(i) Tokenization, sentence extraction and part-of-speech (POS) tagging (i.e. recognizing nouns and verbs)
(ii) Named entity recognition identifies entities such as persons, organizations, locations, numbers (e.g.,time, money amounts). Co-reference detection identifies various spellings or notations of a single named entity.
(iii) Word sense disambiguation identifies the correct sense of a word depending on its context.
(iv) Relationship discovery identifies relations, links and correlations.

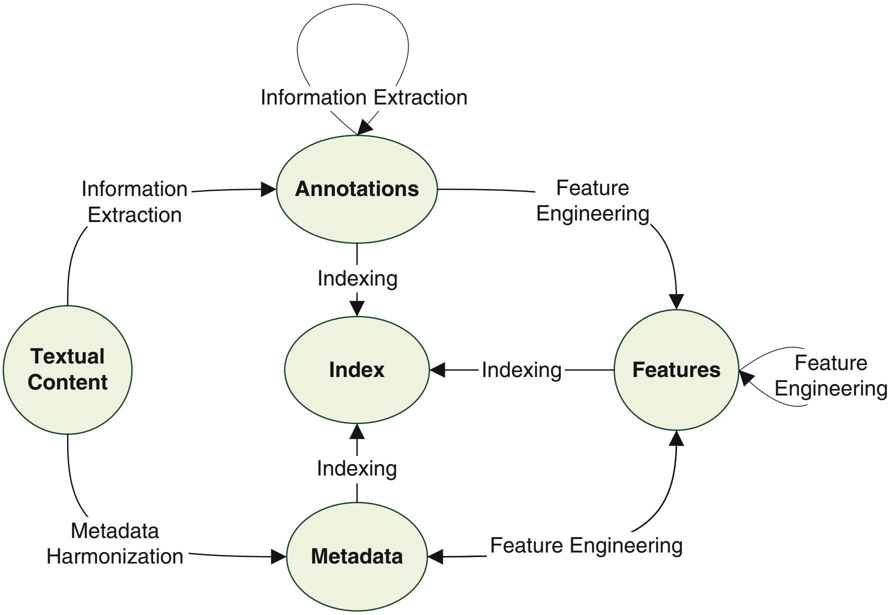

Fig. 7.2 Semantic enrichment steps starting at single artifacts, e.g., documents, news articles, patents (*left*), and resulting in enriched representations in an index (*center*)

7.2.2.2 Feature Engineering and Vectorization

Feature engineering and vectorization uses IE results and document metadata to identify, weight (e.g., TF-IDF), transform (e.g., stemming) and select (e.g., stop-word filtering) features describing text documents. Features are represented as feature vectors used by algorithms to compare documents and compute document similarities. Multiple feature spaces (also referred to as *feature name spaces*) group features of similar characteristic to describe different, potentially orthogonal aspects of a document. For example, one feature space can capture all nouns, while a second feature space can capture all extracted and pre-defined locations and a third all persons. By separating feature spaces, subsequent algorithms can take care of different feature distributions and consider different importance among feature spaces depending on the analytical task at hand, as e.g., in [32]. Besides data integration and cleaning, good feature engineering becomes the second most important step in every Visual Analytics workflow and hence subject for being steered in the analytical process.

Table 7.1 Levels of and techniques used for semantic integration

		Level	
		Schema	Instance
Technique	Declarative-deductive	Shared vocabulary, reasoning-based integration	Shared identifiers (e.g. URIs), rule-based transformation, declarative languages and identifier schemas
	Statistical-inductive	Similarity based on structure, linguistic or data type	Similarity estimates based on entity properties (e.g. clustering, near-duplicate search)

7.2.2.3 Indexing

Indexing develops efficient index structures in order to search for documents containing particular features, or sequences/annotations of features themselves. Inverted indices, representing for each feature the list of occurrences in documents, are among the most often used indexing structures for text. They exploit power law distributions of features in order to allow efficient search and retrieval in text based repositories [34].

7.2.2.4 Text Classification

Text classification employs supervised machine learning methods to organize documents into a predefined set of potentially structured categories [40]. Classification can be seen as injecting structured knowledge via a statistical, inductive process. Examples are assigning documents to topical categories, estimating genre information or determining the sentiment of text passages.

7.2.3 Semantic Integration

Semantic integration aims at integrating information from different, potentially decentralized information sources based on information provided by each source and by previous semantic enrichment processes. With the advancement of decentralized information systems, like the Web, semantic integration becomes a more and more important topic. Semantic integration, also known as *data fusion* in the database community [2], or *ontology mediation* in the Semantic Web community [3], takes place on two levels: the *schema-level* and the *instance-level*. Orthogonal to the two levels, techniques used for integration can be distinguished into *declarative-deductive* and *statistical-inductive* techniques (see Table 7.1 for an overview).

7.2.3.1 Schema Level

The schema level considers mappings of general concepts of objects, like for example mapping the concept *person* in repository A to the concept *people* in repository B. The mapping type relies on the available vocabulary and may range from equality relations to complex part-of relations, depending on the language used. Besides mapping schemas onto each other, schema integration targets the creation of one general schema out of the source schemas. In any case, the result is a shared schema across repositories.

7.2.3.2 Instance Level

The instance level considers mappings of instances of concepts or objects, like for example identifying that person A is the same as person B. While instance-level integration mostly focuses on de-duplication of single instances, more complex cases, like for example determining the type of relationships between two concrete persons, may also be estimated. In general, performance decreases with increasing relationship complexity.

7.2.3.3 Declarative-Deductive Techniques

Declarative-deductive techniques provide mappings based on complex rule sets which may take use of reasoning and/or shared vocabularies. For example, the concept *person* in schema A is the same as *people* in schema B, since both are parents to the disjunctive concepts *man* and *women*. Similarly, on the instance level, the field *name* for the concepts may be the same as the joint field *forename* and *surname* for the concept *people* in schema B. Hence, declarative rules allow to map schemas and instances onto each other. However, declarative rules may not be able to include fuzziness, like different spelling of concept names or aspects like similar structures of concepts/instances.

7.2.3.4 Statistical-Inductive Techniques

Statistical-inductive techniques account the need for fuzzier matching criteria and the capability to learn from example instances. For example, de-duplication of instances – like identifying the set of unique persons from a set of persons – can be solved by clustering instances according to some similarity measures. The identified clusters constitute the unique persons. Similarly, near duplicate search can be used to match a given instance to a set of unique instances, as for example in determining identical web pages during crawling.

Along both dimensions – i.e., levels and techniques – visual analytics can support the integration process by visualizing effects of declarative rules or by visualizing the relation of certain instances to each other. Visual feedback methods allow to steer the integration process. Granitzer et al. [16] present an overview on such visually supported semantic integration processes.

7.2.4 Selection and Aggregation

Semantic enrichment and integration prepare the underlying data set for further processing. In order to reduce the number of documents subject to visual analysis, the next step includes selecting a proper subset and/or aggregating multiple documents to one single object.

7.2.4.1 Retrieval

Retrieval techniques perform the step of selecting appropriate subsets of documents. Besides the capability to search for information in text and metadata, the scalability and performance of modern retrieval techniques [34] enables feature-based filtering, query-by-example and facetted browsing, with the goal to cover all relevant documents for subsequent analysis. Aggregation and visualization methods can be applied to analyze large results sets and drill-down to task specific aspects.

7.2.4.2 Unsupervised Machine Learning

Unsupervised machine learning, in particular clustering, determines groups of similar documents based on the assumption that documents distribute over topics [55]. Groups of documents can be represented as one single data point to reduce the amount of data-points for subsequent processing steps or to improve navigation. Especially for navigation, summarizing clusters in a human readable way becomes crucial. In general, summarizations consist of extracted keywords or a brief textual description relevant to the cluster, created via summarization methods.

7.2.4.3 Summarization

Summarization methods compute a brief descriptive summary for one more documents in the form of representative text fragments or keywords. The summaries are used as labels to represent the essence of a document or a document set. Exploiting integrated metadata and structure between documents becomes essential in order to improve summarization and keyword extraction, as shown in [29]. An overview of different summarization methods can be found in [5].

7.2.5 Visualization and Interaction

The processing steps discussed above return a set of relevant objects, including features, annotations and metadata. As a next step, suitable visual layouts have to be calculated. Text data is characterized by its high-dimensional, sparse representation, which naturally leads to the application of ordination techniques.

7.2.5.1 Ordination

Ordination is a generic term describing methods for creating layouts for high-dimensional objects based on their relationships. It can be seen as a subset of dimensionality reduction techniques [11]. Dimensionality reduction methods project the high-dimensional features into a lower-dimensional visualization space, while trying to preserve the high-dimensional relationships. High-dimensional relationships can be usually expressed by similarity or distance measures. The produced layout is suitable for visualization and exploratory analysis. Other layout generation techniques, such as graph layout methods [6], are used to create a suitable visual layout for non-vector based structures, like typed graphs, temporal processes, etc.

7.2.5.2 Interactive Visualization

Interactive visualizations form the heart of any visual analytics application. Interactive components are used to visually convey information aggregated and extracted from text, and to provide means for exploratory analysis along the lines of the visual analytics mantra: “analyze first – show the important – zoom, filter and analyze further (iteratively) – details on demand” [24]. Feedback provided by users when interacting with visual representations can be fed into the previous stages of the process in order to improve its overall performance.

Visualizations usually depend on the visualized data (e.g., set, tree, graphs) and the task at hand (e.g., topical similarity, temporal development). In Sect. 7.3 we provide a detailed overview on visualizations particularly suited for text.

7.2.6 Hypothesis Formulation and Analytics Workflow

One core difference between Information Visualization and Visual Analytics lies in the support of analytical workflows and the generation and validation of hypothesis. Both, workflows and hypotheses formulation, require support from the underlying analytical process and serve as end-point towards the manipulation of all preceding steps, like acquisition, enrichment, integration, etc.

7.2.6.1 Hypothesis Generation

The visual representation of semantics in the data usually triggers new insights. New insights result in the generation of new, potentially valid hypothesis on the underlying data. For example, showing a distribution of topics in media over time may trigger the hypothesis that two events are related to each other.

Usually hypothesis generation is done in the head of the analyst, rather than making the validated hypotheses explicit. However, hypothesis generation depends on already generated and validated hypothesis. Similar to the well known “Lost in Hyperspace” effect, where users who browse the web via hyperlinks loose their initial information need, implicit hypothesis generation bears the risk to miss important, already validated facts. Hence, hypotheses and the decisions they triggered should be made explicit within an analytical process in order to guide the user. To the best of the authors knowledge this has not been done so far, but well known mathematical models for decision making processes, like the *Analytical Hierarchical Process* (AHP) [36], could be a first starting point therefore.

7.2.6.2 Hypothesis Validation

A generated hypothesis can be verified by the user. Depending on the required manipulation of the underlying analytical process, validation may range from simple interactions with the visual representation to crawling a completely new data set. For example, to see that two events are related to each other one could simply select data points related to both events and reveal their topical dependency. Hence, for efficient support of analytical tasks flexible, powerful and easy to use *task specific workflows* become important.

7.3 Visual Representations and Interactions on Text

Tight integration of visual methods with automatic techniques provides important advantages. To name a few: (i) flexibility to interchangeably apply visual and automatic techniques, as needed by the user, (ii) results of automatic text analysis, such as extracted metadata or aggregated structures, open the way for applying a wider variety of visualization techniques, which are not targeted exclusively to text, and (iii) user feedback can be used to adjust and improve the models used by automatic methods. Therefore, in this section we discuss visual representations, which primarily target textual data, and also describe how visualizations, which do not specifically target text, can be used when information is extracted from text using methods described in the previous section.

7.3.1 Topical Overview

Gaining an overview of important topics in a document set, and understanding the relationships between these topics, is crucial when users are dealing with large text repositories they are unfamiliar with. *Tag clouds* and *information landscape* are examples of visual representations which are designed to address these requirements.

7.3.1.1 Tag Clouds

Tag clouds are a popular Web 2.0 visual representation consisting of terms or short phrases which describe the content of a document or collection. Typically, keywords or named entities (such as persons or organizations) are displayed, which were extracted from document content using natural language processing methods (see Sect. 7.2.2). Size, color and layout of the words are driven by their importance, as well as by aesthetic and usability criteria [44].

In Fig. 7.3 a search result set is visualized by multiple tag clouds combined into one visualization. Each tag cloud corresponds to one of the (pre-defined) categories "sports", "politics", "europe", "society", "culture". The central tag cloud represents all documents of all categories. Each tag cloud shows the most important named entities (persons, dates, locations) for the respective category, thus giving an overview over the documents within. The polygonal boundaries for each tag cloud are generated by applying Voronoi subdivision. The initial points for generating this subdivision can either be set manually (as in the example figure) or can be the result of a similarity layout of the category content (for an example, see [46]).

7.3.1.2 Information Landscapes

Information landscapes, such as In-SPIRE [27] and InfoSky [1], employ a geographic landscape metaphor for topical analysis of large document sets. Information landscapes are primarily used for gaining an overview and for providing explorative navigation possibilities. A user who is unfamiliar with the data set is empowered to gain insight deep into the topical structure of the data, understand importance of various topics, and learn about relationships between them. As opposed to searching using queries, guided explorative navigation provides the possibility to identify interesting information even when the user's goals are vaguely defined.

In information landscapes documents are visualized as dots (icons), which are laid out in such a way that similar items are positioned close together, while dissimilar ones are placed far apart. Hills emerge where density of topically related documents is high, indicating a topical cluster. Clusters are labeled by summaries of the underlying documents, allowing users to identify areas of interest and eliminate outliers. The height of a hill is an indicator for the number of documents and the

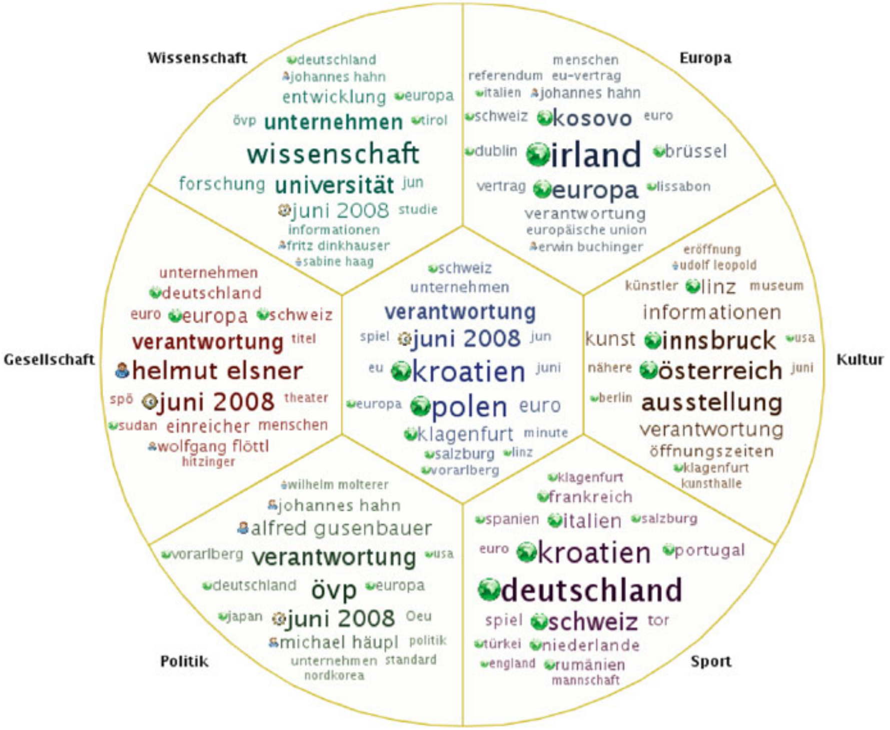

Fig. 7.3 A visualization showing a search result set as a combination of tag clouds. Each polygonal area corresponds to a category of the documents in the search result set. Displayed named entities are enhanced with symbols indicating their type (person, location, data)

compactness of the hill is an indicator of cluster's topical cohesion. Topically similar clusters can be identified as they will appear spatially close to each other, while dissimilar clusters are separated by larger areas, visualized as sea. Aggregation of the data set and its projection into the 2D space are computed using scalable clustering and ordination algorithms, as for example described in [30, 38] (also see Sects. 7.2.4 and 7.2.5). Advanced information landscape implementations can handle data sets with far over a million documents. For such massive data sets information retrieval techniques (see Sect. 7.2.4) can be used to provide fast filtering and highlighting functionality.

Figure 7.4 shows navigation in an information landscape along a hierarchy of topical clusters, which are visualized as nested polygonal areas. Cluster labels provide a summary of the content of the underlaying documents and serve as guidance for exploration. Following the labels on each level of the hierarchy, the user can navigate the topical structure of the data and understand how clusters relate in terms of topical similarity and size. On the top-left of the figure, an overview of approximately 6,000 news articles on "computer industry" can be seen,

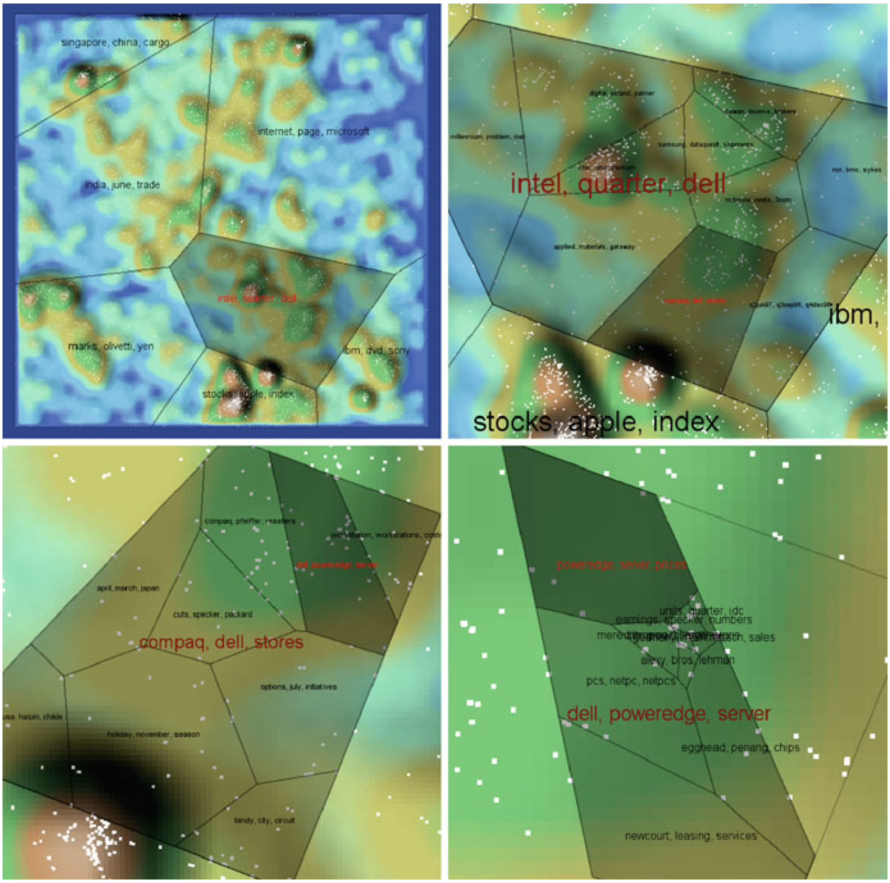

Fig. 7.4 An information landscape showing approx. 6,000 news articles on "computer industry" is used for drilling down to documents of interest: beginning with an overview (*left*) the user narrows down using topical cluster labels (*right*)

subdivided into 7 topical clusters. Clicking on the label "intel, quarter, dell", the corresponding cluster is zoomed in and the sub-areas, corresponding to its sub-clusters, are shown (top-right). Clicking on "compaq, dell, stores" (bottom-left) and then on "dell, poweredge, server" (bottom-right) narrows further down to the potential topic of interest. The cluster "poweredge, server, prices" (bottom-right) contains only five document, which can be inspected manually by the user. Free navigation by zooming (mouse-wheel) and panning (mouse-drag) is also available. Selection of documents can be performed cluster-wise, individually or on arbitrary subsets using a lasso tool.

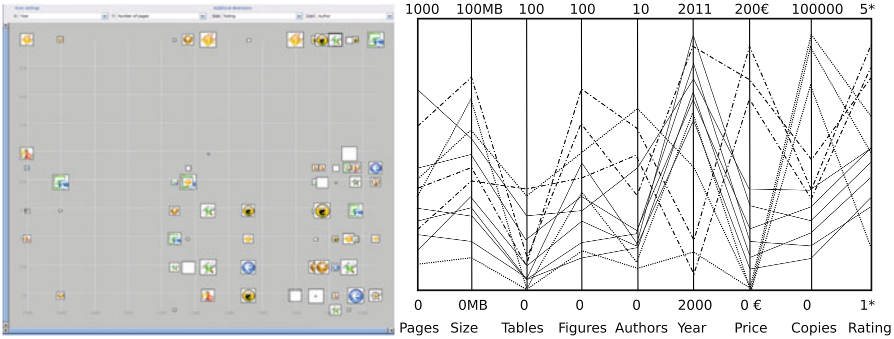

Fig. 7.5 Multidimensional visualization for books. *Left*: Scatterplot visualizing publication year (x-axis), page count (y-axis), file size (icon size), author (icon type); *right*: parallel coordinates showing nine metadata types on parallel axes

7.3.2 Multidimensional Metadata

Visualization of multidimensional metadata enables the discovery of correlations between document metadata. Such metadata may include document size and source, relevance to a search query (see Sect. 7.2.4), or extracted persons and organizations (see Sect. 7.2.2).

7.3.2.1 Scatterplot

Scatterplot is a visual representation for analysis of multidimensional metadata, mapping up to five different metadata types (dimensions) to the x and y axes, and to visual properties (color, size, shape) of displayed items [21]. The main drawback of a scatterplot is that it can correlate only a limited amount of dimensions.

7.3.2.2 Parallel Coordinates

The parallel coordinates representation [19] can handle a larger amount of dimensions, which are displayed as parallel vertical axes. For each document, the variables are displayed on their corresponding axes and connected with a polygonal line, so that patterns can be spotted easily as lines having similar shapes. In addition, a selected discrete property (e.g., class membership) can be mapped to the line color to allow identification of differentiating features for different values of the property.

Figure 7.5 (left) shows a scatterplot displaying book metadata (publication year, page count, file size, author). The scatterplot component builds upon the

Prefuse Information Visualization Toolkit[1] adding the capability to handle multiple coordinated scatterplot views (see Sect. 7.3.6 for more information on multiple coordinated views). By visualizing the same data set in two or more coordinated scatterplots at the same time, it becomes possible to increase the number of visualized dimensions above the typical five. A parallel coordinates visualization in Fig. 7.5 (right), shows nine different types of metadata for e-books, with the line style differentiating between publishers. It can be seen that the some e-books have high ratings and high prices (dashed-dotted lines), some others are cheaper and have lower ratings (continuous lines), while the remaining e-books are free and achieve highest delivery rates (dotted lines).

7.3.3 Space and Time

Visualization of geo-spatial and temporal information is very important in many applications. In what follows, we explain different approaches for producing such visualizations and discuss ThemeRiver [17], a well-known visualization conveying topical changes in large text repositories.

7.3.3.1 Visualization of Geo-Spatial Information

The visualization of geo-spatial information, as for example extracted locations, is a natural fit for the application of various geo-visualization approaches [7]. A popular application of geo-visualization is to show automatically extracted spatial information (see Sect. 7.2.2) on geographical maps [39] in order to reveal where something is happening. Figure 7.6 shows a geo-spatial visualization of locations extracted from German news articles [28]. The extracted locations are depicted on a map of Austria as cones, where the size of a cone corresponds to the number of news articles the location occurred in. Clicking on a cone triggers a filtering of the news article set by the selected location, and thus this visualization can be used as a faceted search tool.

Geo-spatial visualizations are not restricted to geographic maps; they can also be applied in e.g., virtual 3D environments. An example is the planetarium that has been integrated into an encyclopedia application [26], providing coordination between browsing spatial (astronomic) references in text and navigation in the virtual environment.

[1] http://prefuse.org/.

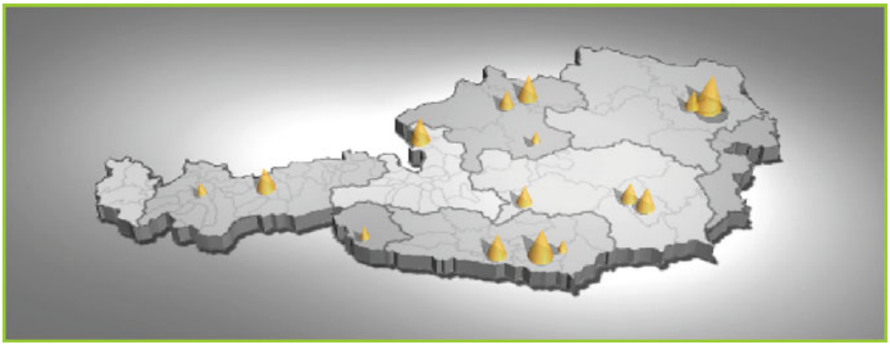

Fig. 7.6 Geo-visualization of Austria showing geo-references in news articles (*cones*). The size of the cone corresponds to the number of news articles for the particular geo-reference

7.3.3.2 Visualization of Temporal Information

The visualization of temporal information, such as document creation date or automatically extracted time references (see Sect. 7.2.2), can be realized by a variety of visual components. Although different in many aspects, visual representations for temporal data usually share a common feature: they include a visual element which symbolizes the flow of time. For example, temporal data can be visualized along a straight line or along a spiral [54], both representing the flow of time. Although a straight time axis is more common, a spiral time axis has the advantage of being suitable for detecting cycles and recurring events, and it allows for displaying long time intervals with high temporal resolution even on small screens.

7.3.3.3 ThemeRiver

ThemeRiver [17] is a well-known visualization conveying topical changes in large text repositories. It uses a metaphor of flowing river streams to visualize trends and changes in topical clusters, in the context of external events (see clustering in Sect. 7.2.4). In addition to topical clusters, metadata clusters, for example documents mentioning a specific location, can also be visualized. ThemeRiver empowers users not only to understand trends but also to discover correlations and causal relationships between clusters.

In Fig. 7.7 a stream visualization, which closely resembles the ThemeRiver, shows temporal development of topical clusters for approximately 750 news documents on “oil spill”. The x-axis symbolizes the flow of time, while the y-axis conveys the amount of documents at a given moment in time. Each topical cluster is represented by a stream of particular color, where the width of the stream along the time axis correlates with the number of documents. By observing the development of the “japan, tokyo, bay” topical cluster (second from bottom), which has two distinctive peaks, it is obvious that temporal development of the “russia” metadata cluster (bottom-most) correlates with the first peak, but not with the second.

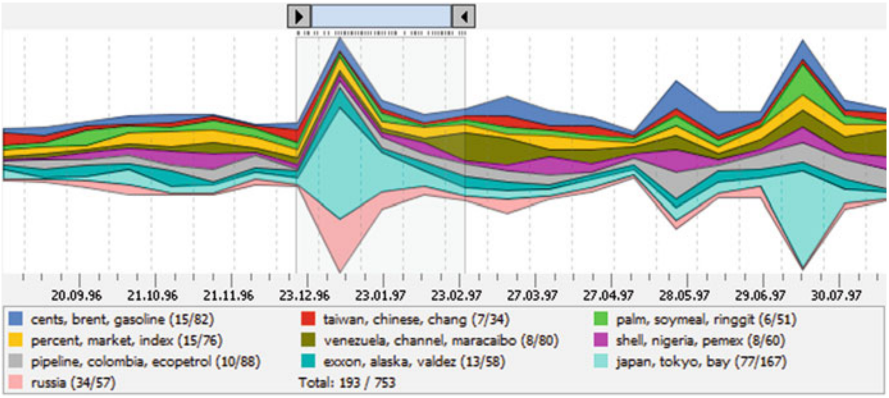

Fig. 7.7 A stream visualization of approx. 750 news documents on "oil spill", showing temporal development. Different gray values correspond to different topics

Naturally, a fusion of both spatial and temporal information in one visualization also leads to interesting results. For example, the three-dimensional GeoTime [22] visualization depicts a geographic map where the flow of time is orthogonal to the map (i.e. on the z-axis). In this way GeoTime facilitates tracking of ground movements over time and identification of activity hot-spots in both space and time.

7.3.4 Relationships

Relationships between concepts (e.g., keywords or named entities), identified by methods such as co-occurrence analysis and disambiguation techniques (see Sect. 7.2.3), are typically presented using graph visualizations [18]. For example, PhraseNet [53] displays relationships between terms within a document, while FacetAtlas [4] relies on faceted retrieval to visualize relationships between faceted metadata. Relationships between aggregated structures (see Sect. 7.2.4), such as document clusters, can be visualized by Cluster Maps [10]. It is a representation similar to Venn and Euler diagrams, showing whether (and through which features) different clusters overlap topically.

7.3.4.1 Graph Visualization

A graph visualization that is used to present relationships extracted from approximately 25,000 documents can be seen in Fig. 7.8. Concepts (keywords) are placed in the 2D plane, depending on their interconnectedness, using a force-directed placement method (see Sect. 7.2.5). An edge bundling technique [25] is applied to reduce clutter, which would otherwise occur due to the high number of relationships.

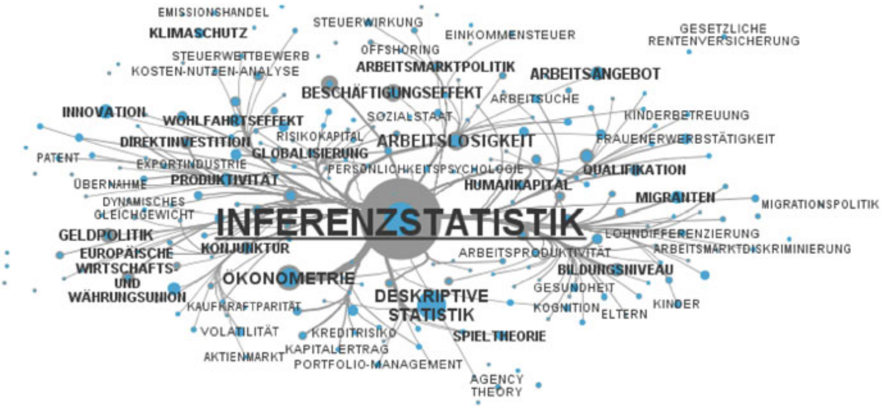

Fig. 7.8 A graph visualization of relationships between concepts extracted from a text data set (data courtesy of German National Library of Economics, 2011). Note that edge bundling is used to improve clarity and reduce clutter in the edge layout

To preserve clarity even when visualizing larger graphs, a level-of-detail sensitive algorithm decides which informations is displayed and which is hidden depending on user focus and the current zoom level. To navigate, the user clicks on a concept which triggers a zoom-in operation focusing that concept. Concepts close to the chosen one are displayed in more details, revealing finer structures in the graph.

7.3.5 Visually Enhanced User Feedback

Analytical tasks require well-designed interaction mechanisms along with different kinds of visualizations. Interactions can be grouped along three orthogonal dimensions, namely (i) the kind of *operation* they perform, i.e., navigation, selection and manipulation, (ii) the *modality of the interaction*, i.e., query, point-and-click, language input, multi-touch, and (iii) its *influence on the underlying analytical process*, i.e., the adaptation of data, parameters or the mining models themselves. This subsection will briefly discuss (i) and (ii), and then focus on how interactions can be used to steer the underlying analytical process.

7.3.5.1 Modalities of Interaction

Modalities of interaction depend mostly on the input devices. A search box can be seen as "textual modality" which allows to filter relevant documents based on keywords, provided either via keyboard or speech-to-text. Clearly, with the advance of multi-touch devices new capabilities in expressing user needs become available. While modalities of interaction influence the design of visualization and determine how interactions take place, they do not influence the possible operations on the data and the steering capabilities on the analytics process. For a detailed overview of different interaction modalities the reader is referred to [49].

7.3.5.2 Operations of Interactions

Operations of interactions describe the purpose of an interaction. Interactions to navigate complex information spaces, to drill down on particular interesting patterns and to switch between different perspectives, are the most common navigational operations. Examples are browsing hyper-links or navigating a hierarchical structure (see Fig. 7.4, left). Selection comprises operations that allow users to select data points of interest and their properties. Examples include multi-selection in a list of documents or lasso selection of data points in a similarity layout of documents (see Fig. 7.4, top-right). Finally, manipulations form the essence of any visual analytics application. Manipulative interactions, like removing certain data points from the analysis or assigning a group of documents to a particular class, allow to steer the underlying classification, clustering and retrieval processes.

7.3.5.3 Steering the Visual Analytics Process

Given interactions of different modalities and operations, the question remains how the underlying process could be steered. In the following, we will discuss steering on the *data-point level* and the *model level*.

Data-point level: On the data-point level, selecting a particular subset of data points or a subset of data sources for the detailed analysis becomes the most common form of influence. For example, Fig. 7.9 (left) shows a similarity layout comparing search results from different search engines [51]. Sources could be interactively added or removed in order to change the topical layout. Similarly, in the landscape visualization shown in Fig. 7.4, a lasso selection can be used to select a set of similar documents. This set could then be used as a positive set of examples for training supervised classification algorithm.

Model level: On the model level, the goal is to control the underlying mining model. The most direct form of controlling mining models is by setting parameters directly, for example, the number of clusters or cost functions for positive or negative classification errors. However, comprehension of resulting effects of direct parameter manipulation becomes non-trivial especially for data-mining laymen. Hence, we propose "direct manipulations" of mining models using visualizations.

The concept of direct manipulation greatly improved user interfaces of computers by allowing users to directly manipulate information objects, like files and folders.

Users have been empowered to drag and drop object instead of manipulating them via a command line. In analogy, we give two examples on direct manipulation in visual analytics.

Example 1. As a first example, consider again the landscape visualization shown in Fig. 7.4. High-dimensional data points have been projected onto the 2D-plane using clustering and ordination techniques. However, the high-dimensional distance measure, and therefore the result of the projection, may not fit to the users expectation of "distance" between documents. Some topics may be too close and some too far away from each other. Instead of directly changing the distance function or parameters of the ordination technique, the user could directly drag topically similar data points closer to each other and dissimilar data-points further apart, yielding user-determined distances between data points. By applying metric learning techniques [14, 47] the user-determined similarity could be transformed into a high-dimensional distance function in some kind of inverse projection [15]. The resulting high-dimensional distance function can be used in different mining algorithms to reflect the user's notion of "similarity" between documents.

Example 2. A second example concerns supervised machine learning models. Interactions on a visual layout may be used two-fold: (i) to correct classification errors or re-force correct classifications, and (ii) to generate new training data. A visualization supporting these tasks requires the following properties: First, the visualization should allow to judge problematic behavior of classification models, like biases towards particular classes. Second, fast and easy identification of false and/or problematic examples, e.g., outliers should be supported. Third, users should be able to rapidly select and (re-)label examples. Further, if is preferable to have the same kind of visualization, independently of the classification task and the employed data classification algorithm.

The visualization proposed in [42, 43] satisfies these properties and has been shown to support users in improving classification models [41]. Here, classes are arranged around a circle. Data-points are placed in the interior of the circle with their distances to every class being proportional to the a-posterior probability that a data point belongs to that class (see Fig. 7.9, right). Data points can be inspected, selected and dragged to the correct class resulting in re-training and improving the underlying text classifier. A combined user interface employing this visualization and an information landscape has further been shown applicable to generate classifier hypothesis from scratch [45].

7.3.6 Multiple Visualization Interfaces

Complex analytic scenarios involve heterogeneous data repositories consisting of different types of information. Visual representations are designed to target specific aspects of the data, such as metadata correlations, topical similarity, temporal

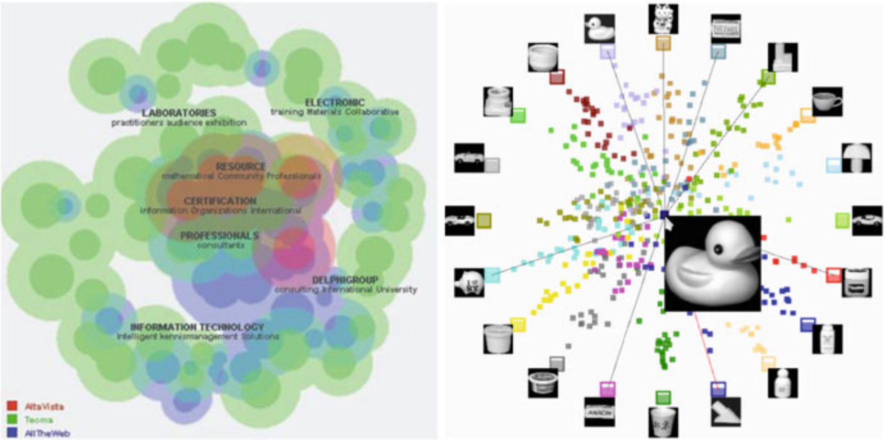

Fig. 7.9 Examples for visually enhanced feedback. *Left*: Search results (*circles*) for comparing the topic overlap of different search engines (colors). Results with similar content are close. Sources can be interactively added or removed. *Right*: Visualizing classification decisions. Classes are arranged in a circle, data points are placed inside the circle according to their a-posteriori probabilities. Decisions can be corrected by drag and drop, classifier is retrained

developments, geo-locations, etc. When simultaneous analysis of different information types is required, user interfaces consisting of multiple visual components are necessary. One way to address visual analysis of heterogeneous data is to integrate various visualizations within a single immersive 3D virtual environment, such as the Starlight System [35]. A more widely used approach is *Coordinated Multiple Views* (CMV) [31]. Multiple view coordination is a technique for tightly coupling multiple visualization components into a single coherent user interface, so that changes triggered by interactions in one component are immediately reflected in all others components.

Figure 7.10 shows a coordinated user interface, consisting of an information landscape, a stream visualization, as well as of several other widgets, such as trees and tables. The interface is used for "fused" analysis of topical, temporal and metadata aspects of large text repositories [37]. The tree component, on the left, shows the hierarchy of topical clusters providing a virtual table of contents. An information landscape (see Sect. 7.3.1), on the right, visualizes document frequency and topical similarity of clusters and documents. A stream view (see Sect. 7.3.3), on the bottom, conveys temporal development of topical (and metadata) clusters. Two additional components are available but are hidden in the screenshot: a faceted metadata tree showing extracted persons, organizations and locations, and a table providing detailed information on clusters and documents.

The coordination of components includes the following: (i) *navigation in the cluster hierarchy* (triggered in any of the components), (ii) *document selection*

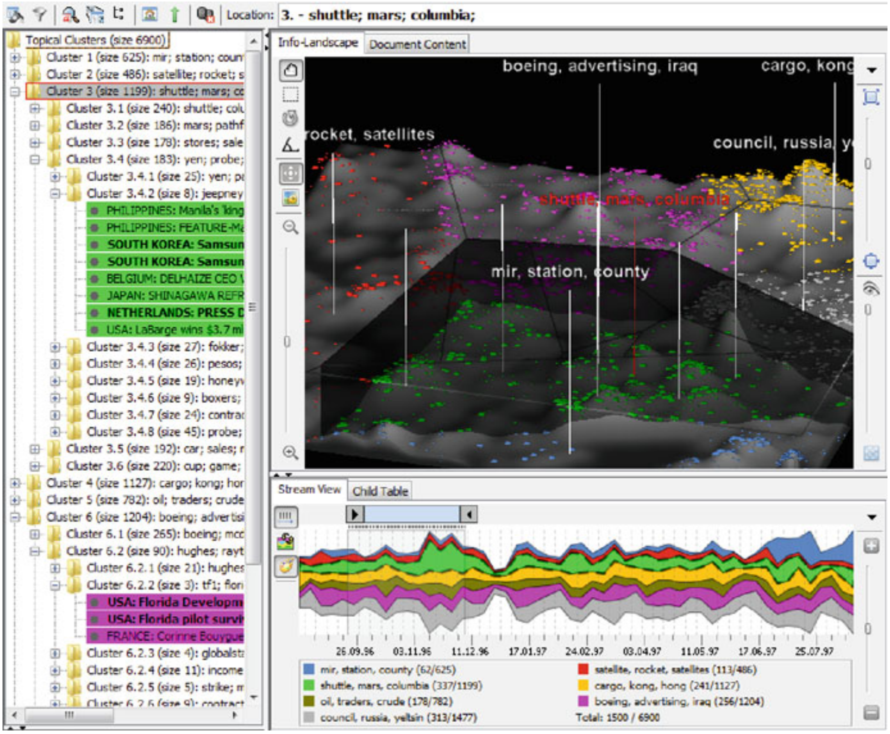
Location: 3. - shuttle; mars; columbia;
Topical Clusters (size 6900)
Cluster 1 (size 625): mir; station; coun
Cluster 2 (size 486): satellite; rocket; s
Cluster 3 (size 1199): shuttle; mars; co
Cluster 3.1 (size 240): shuttle; colu
Cluster 3.2 (size 186): mars; pathf
Cluster 3.3 (size 178): stores; sale
Cluster 3.4 (size 183): yen; probe;
Cluster 3.4.1 (size 25): yen; pe
Cluster 3.4.2 (size 8): jeepney
PHILIPPINES: Manila's 'king
PHILIPPINES: FEATURE-Ma
SOUTH KOREA: Samsun
SOUTH KOREA: Samsun
BELGIUM: DELHAIZE CEO
JAPAN: SHINAGAWA REFR
NETHERLANDS: PRESS D
USA: LaBarge wins $3.7 mi
Cluster 3.4.3 (size 27): fokker;
Cluster 3.4.4 (size 26): pesos;
Cluster 3.4.5 (size 19): honeyw
Cluster 3.4.6 (size 9): boxers;
Cluster 3.4.7 (size 24): contrac
Cluster 3.4.8 (size 45): probe;
Cluster 3.5 (size 192): car; sales; r
Cluster 3.6 (size 220): cup; game;
Cluster 4 (size 1127): cargo; kong; hor
Cluster 5 (size 782): oil; traders; crude
Cluster 6 (size 1204): boeing; advertisi
Cluster 6.1 (size 265): boeing; mcc
Cluster 6.2 (size 90): hughes; rayt
Cluster 6.2.1 (size 21): hughes
Cluster 6.2.2 (size 3): tf1; flori
USA: Florida Developm
USA: Florida pilot surviv
FRANCE: Corinne Bouygue
Cluster 6.2.3 (size 4): globalst
Cluster 6.2.4 (size 11): income
Cluster 6.2.5 (size 5): strike; m
Cluster 6.2.6 (size 9): contract
Info-Landscape
Document Content
boeing, advertising, iraq
cargo, kong
rocket, satellites
council, russia, y
shuttle, mars, columbia
mir, station, county
Stream View
Child Table
26.09.96
03.11.96
11.12.96
17.01.97
24.02.97
03.04.97
11.05.97
17.06.97
25.07.97
mir, station, county (62/625)
satellite, rocket, satellites (113/486)
shuttle, mars, columbia (337/1199)
cargo, kong, hong (241/1127)
oil, traders, crude (178/782)
boeing, advertising, iraq (256/1204)
council, russia, yeltsin (313/1477)
Total: 1500 / 6900

Fig. 7.10 A coordinated multiple views GUI showing 6,900 news documents on "space". Document selection (by time: from June to August), document coloring (each topical cluster in different color) and navigation in the hierarchy (location: Cluster 3 "shuttle, mars, columbia") are coordinated

(lasso-selection in the landscape, temporal selection in stream view, or cluster-wise selection in the trees), (iii) *document coloring* (driven by the stream view color assignments), and (iv) *document icons* (user-assignable from any component).

Coordination ensures that all views will focus on the same cluster, and that document selection, colors and icons are consistent in all views. In this way, discovery of patterns over the boundaries of individual visualizations becomes possible. For example, topical-temporal analysis can be performed by selecting documents belonging to two temporally separate events in the stream view, and then inspecting in the landscape whether those documents are topically related or not. Moreover, correlations between topical clusters and occurrences of a metadatum (e.g., persons) can be identified by assigning different icons to documents mentioning different persons, and then observing the distribution of these persons over topical clusters in the landscape.

7.4 Application Scenarios and Domains

Application scenarios for visual analysis and discovery in text repositories can be identified in a wide range of domains. News media, encyclopedia volumes, scientific paper repositories, patent databases or intelligence information systems, represent an exemplary selection of domains to which the methods discussed in this chapter have been beneficially applied. Given the diversity of application scenarios, we will try to impose some structuring along relevant dimensions.

A first important dimension involves the target user group of a given application. Clearly, the skill and experience level of the expected target user group should influence the choice of visual means. *Information Visualization* and *Visual Analytics* approaches usually focus on efficiency. This results in visual means which are perfectly suitable for expert analysts, who have a high level of visual literacy and domain knowledge. In contrast, *Knowledge Visualization* approaches [8] focus on comprehensibility. The resulting visual means are often less efficient and flexible, but can be utilized by a general audience.

A second dimension considers the amount of a priori information and context available in a given application scenario. If information or context is available, for example in the form of a formulated query or user profile information, an initial search can limit the number of information items which have to be considered. In this case, the visual analysis and manipulation of search results becomes the prevalent task. In the absence of explicit information or context, explorative visualizations can enable the discovery of facts without having to explicate an information need in advance. The following application scenarios provide a representative cross section along this dimensions.

7.4.1 Media Analysis for the General Public

Media Analysis providers have traditionally shaped their services towards the requirements of decision makers in enterprises and organizations. The advent of the World Wide Web and the introduction of consumer-generated media has greatly increased the amount of news sources available to a general audience. Media consumers today find themselves assuming the role of media analysts in order to satisfy personal information needs. News visualization has been a favored use case for Information Visualization almost from the beginnings of this discipline [33]. However, in the spirit of the structure established above, visual support for this application scenario should employ simple visual means and assume limited visual literacy.

The *Austrian Press Agency* (APA) has provided a general audience with a number of experimental news visualizations through its labs platform since 2008 [28]. From a technical point of view, the platform implements the pipeline architecture outlined in this chapter. The acquisition stage relies on the PowerSearch media

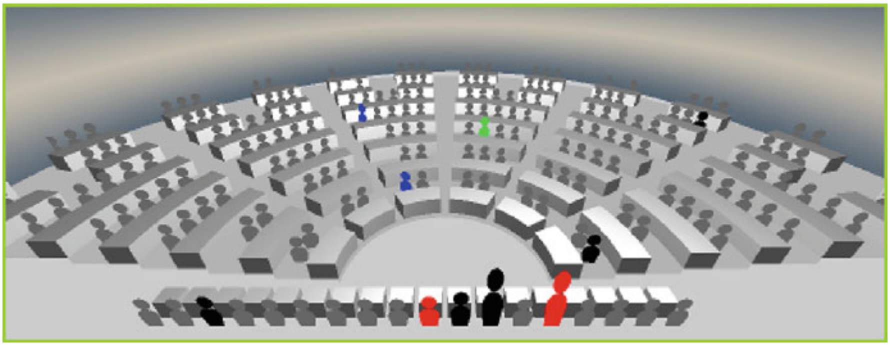

Fig. 7.11 A visualization of occurrences of Austrian politicians in search results. A rendered model of the parliament is used as visual metaphor. The figures of politicians are colored in their party color and scaled relative to the occurrence count. Clicking on a figure narrows the search result to articles containing the selected politician

database run by the company, which provides 180 Million news articles from 250 sources in a normalized manner. Semantic enrichment is facilitated through a combination of rule-based and dictionary-based methods, which annotate persons, locations and companies. Machine learning techniques are used to classify articles into topical areas. Semantic integration is currently being addressed, for instance by harmonizing identified persons with appropriate data sources from encyclopedias. Retrieval is performed through a classical query-based interface, which provides relevance-ranked search result lists.

The initial architecture has been tailored towards faceted filtering of large search result sets. Given a query entered by a user, the system generates the result set and displays a variety of visualizations, each of which represents a certain facet. For instance, the occurrence of members of parliament and members of government in a set of search results is visualized in a model of the Austrian parliament, as shown in Fig. 7.11. Other visualizations include a geo-spatial view, a round table view of prominent politicians and a tag cloud. All visualizations are very simple in design, rely on metaphors to ease understanding and support a very simple interaction scheme: Selecting a visual entity filters the result set to results containing the entity. Experiments have shown that this kind of system is accessible to a general audience without training.

An example visualization of more complex media analysis results is shown in Fig. 7.12. Co-occurrence of key political figures extracted from a text corpus is represented using a node-link-diagram in which links have been bundled to reveal high-level patterns [25]. This kind of visualization favors an exploratory approach which reveals general trends of the whole article set in the absence of a concrete search query.

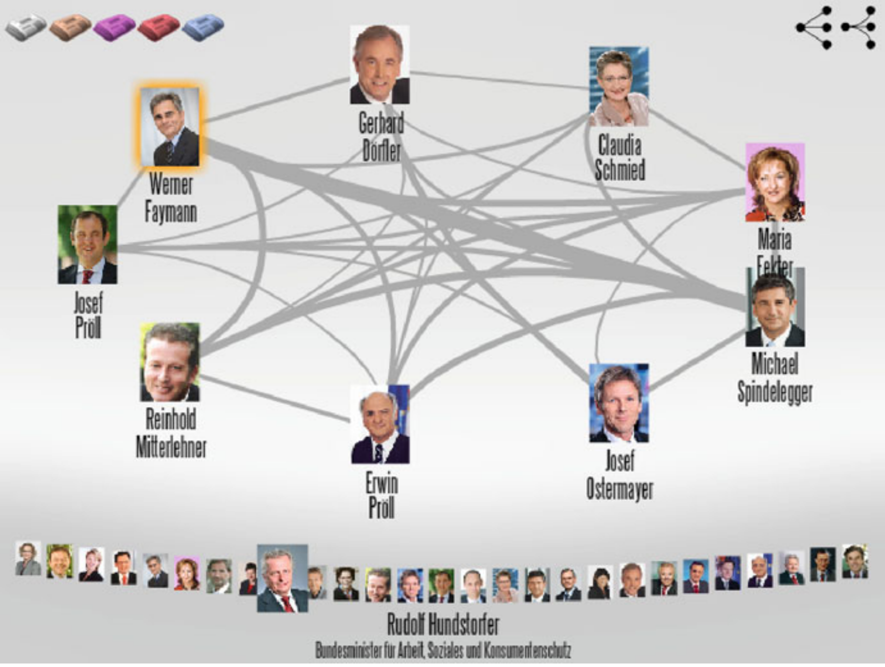

Fig. 7.12 A visualization of co-occurrences of Austrian politicians in recent news media. Politicians are displayed as nodes connected by links representing co-occurrence strength by line width. Links are bundled to reveal high-level edge patterns. The strongest link visible is between the chancellor and the vice-chancellor

7.4.2 Navigation and Exploration of Encyclopedias

Modern digital encyclopedias contain hundreds of thousands of textual articles and multimedia elements, which constitute a knowledge space encompassing virtually all areas of general interest. Traditional retrieval and discovery techniques in this domain have included keyword search for articles and cross-reference based navigation between articles. The German-language Brockhaus encyclopedia provides a visualization system which enables the visual navigation of article context. This three-dimensional Knowledge Space visualization presents topically related articles, using figurative graphical elements as visual metaphors. The idea behind the visualization is to support navigation between articles and to encourage exploration of the encyclopedia in the spirit of edutainment.

The visualization shown in Fig. 7.13, displays the currently selected article at the center of a disc divided into topical segments and arranges similar articles around it. Relevant articles are placed close to the center and each article is placed within the segment corresponding to its topic (chosen from a ten-item topic scheme). Articles are represented by shapes according to type: *cylinders* represent

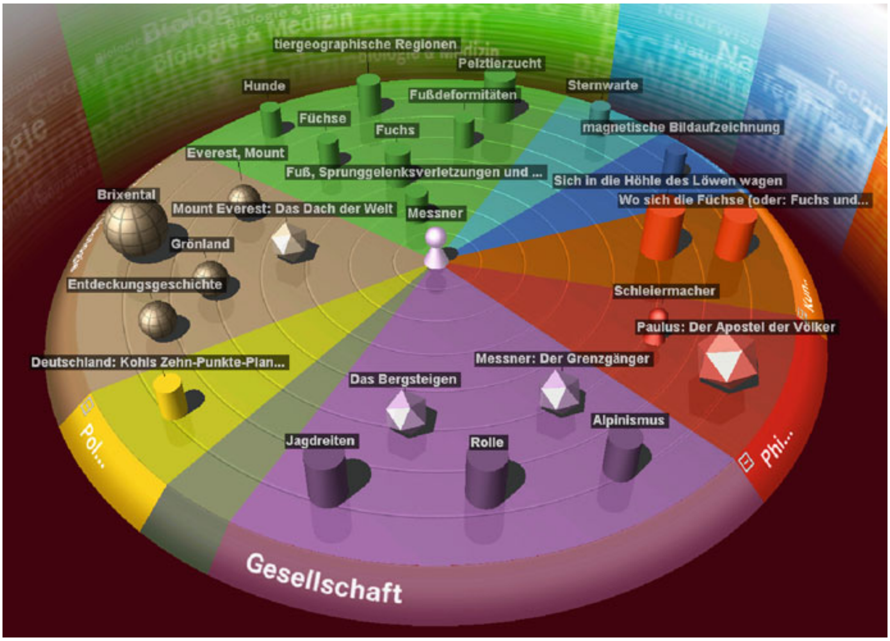
Biologie & Medizin
tiergeographische Regionen
Pelztierzucht
Hunde
Fußdeformitäten
Sternwarte
Füchse
Fuchs
magnetische Bildaufzeichnung
Everest, Mount
Fuß, Sprunggelenksverletzungen und ...
Sich in die Höhle des Löwen wagen
Brixental
Mount Everest: Das Dach der Welt
Messner
Wo sich die Füchse (oder: Fuchs und...
Grönland
Entdeckungsgeschichte
Schleiermacher
Paulus: Der Apostel der Völker
Deutschland: Kohls Zehn-Punkte-Plan...
Messner: Der Grenzgänger
Das Bergsteigen
Alpinismus
Jagdreiten
Rolle
Pol...
Phi...
Gesellschaft

Fig. 7.13 The "Knowledge Space" visualization displaying the context of the encyclopedia entry for the mountaineer Reinhold Messner (*center*). The disc is divided into segments representing topics (e.g., "society" in the front). Related articles are represented by objects placed on the disc; shape, size and color encode additional metadata. For example, in the leftmost segment a geographic article (*circle*) and a premium content article (*diamond*) about the Mountain Everest is shown

factual articles, *spheres* represent geographic articles, *cones* represent biographic articles and *diamonds* represent articles featuring premium content. Article labels are displayed above the shapes. Dragging the mouse horizontally spins the disc around its central axis. Dragging the mouse vertically adjusts zoom factor and vertical view angle. Clicking on an object navigates to the corresponding article.

7.4.3 Patent Analysis and Comparison

The identification of prior art, and the discovery of patterns and trends in patents constitutes a crucial aspect of business intelligence for innovative enterprises. The raw data for this kind of analysis is readily available in the form of various commercial and open patent databases. However, the actual information contained in patents is very hard to analyze and understand. This phenomenon stems, in part, from deliberate attempts to paraphrase key issues in order to maintain a competitive advantage. Another reason for the complexity of patent information is the huge amount of domain knowledge required to make sense of an actual patent, covering a narrow technical aspect.

The Austrian company *m2n* has created a patent analysis system which has been used by various large enterprises, for instance by one of the largest global steel manufacturers. This system displays patent data sets, acquired from a number of configurable sources, in a multiple coordinated view environment, which integrates textual and visual representations [37]. The visualization application includes an information landscape, a temporal visualization and a number of other coordinated views, similar to the user interface shown in Fig. 7.10. Referring to the structure established above, this system clearly targets expert users which accept a large amount of training in order to harvest all the benefits.

7.5 Conclusion and Outlook

Through combining visually supported reasoning with large scale automatic processing, visual analytics opens new possibilities for exploration and discovery of knowledge in text repositories. Aggregation and summarization are central to scaling visualizations to very large data sets. Retrieval techniques enable filtering, highlighting and selection on repositories of virtually unlimited size. Information extraction opens the way for using visual representations which are not directly related to text, such as geo-visualization or graph visualization. Finally, visualization not only introduces human knowledge and visual pattern recognition into the analytical process, but also provides the possibility to improve the performance of automatic methods through consideration of user feedback.

While it is hard to deliver predictions on future development of the field, the following directions appear promising: Triggered by the surge in use of smart mobile devices and multi-touch interfaces, support for collaborative scenarios using new input devices, such as tablets and multi-touch tables, is likely to gain traction. On the algorithm side, the peculiarities of the emerging phenomenon of social networks and social media, such as quality and trustworthiness of information, pose new challenges. In the quest to handle ever larger data sets the efficient exploitation of the cloud for computation and storage holds the promise of ultimate scalability.

References

1. Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., Tochtermann, K.: The infoSky visual explorer: exploiting hierarchical structure and document similarities. Inf. Vis. **1**(3–4), 166–181 (2002)
2. Bleiholder, J., Naumann, F.: Data fusion. ACM Comput. Surv. **41**, 1:1–1:41 (2009)
3. Bruijn, J.d., Ehrig, M., Feier, C., Martìns-Recuerda, F., Scharffe, F., Weiten, M.: Ontology Mediation, Merging, and Aligning, in Semantic Web Technologies: Trends and Research in Ontology-based Systems (eds J. Davies, R. Studer and P. Warren), John Wiley & Sons, Ltd, Chichester, UK. pp. 95–113. (2006). doi:10.1002/047003033X.ch6

4. Cao, N., Sun, J., Lin, Y.R., Gotz, D., Liu, S., Qu, H.: Facetatlas: multifaceted visualization for rich text corpora. IEEE Trans. Vis. Comput. Graph. **16**(6), 1172–1181 (2010)
5. Das, D., Martins, A.F.: A survey on automatic text summarization. Technical report, Carnegie Mellon University (2007). Literature Survey for the Language and Statistics II course at CMU
6. Díaz, J., Petit, J., Serna, M.: A survey of graph layout problems. ACM Comput. Surv. **34**, 313–356 (2002)
7. Dykes, J., MacEachren, A.M., Kraak, M.J. (eds.): Exploring Geovisualization. Elsevier, Amsterdam (2005)
8. Eppler, M.J., Burkhard, R.A.: Knowledge visualization. In: Schwartz, D. & D. Te'eni (eds.) Encyclopedia of Knowledge Management, Second Edition, PA: Information Science Reference. pp. 987–999. Hershey. doi:10.4018/978-1-59904-931-1.ch094
9. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**, 37–54 (1996)
10. Fluit, C.: Autofocus: semantic search for the desktop. Inf. Vis. Int. Conf. **0**, 480–487 (2005)
11. Fodor, I.: A survey of dimension reduction techniques. Technical report UCRL-ID-148494, US DOE Office of Scientific and Technical Information (2002)
12. Gantz, J.F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., Manfrediz, A.: The expanding digital universe, a forecast of worldwide information growth through 2010. IDC White Paper – sponsored by EMC (2007)
13. Gantz, J.F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., Toncheva, A.: The diverse and exploding digital universe, an updated forecast of worldwide information growth through 2011. IDC White Paper – sponsored by EMC (2008)
14. Granitzer, M.: Adaptive term weighting through stochastic optimization. In: Gelbukh, A. (ed.) Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, vol. 6008, pp. 614–626. Springer, Berlin/Heidelberg (2010)
15. Granitzer, M., Neidhart, T., Lux, M.: Learning term spaces based on visual feedback. In: International Workshop on Database and Expert Systems Applications (DEXA), Krakow, pp. 176–180. IEEE Computer Society (2006)
16. Granitzer, M., Sabol, V., Onn, K.W., Lukose, D., Tochtermann, K.: Ontology alignment – a survey with focus on visually supported semi-automatic techniques. Future Internet **2**(3), 238–258 (2010)
17. Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: visualizing thematic changes in large document collections. IEEE Trans. Vis. Comput. Graph. **8**(1), 9–20 (2002)
18. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. IEEE Trans. Vis. Comput. Graph. **6**, 24–43 (2000)
19. Inselberg, A., Dimsdale, B.: Parallel coordinates for visualizing multi-dimensional geometry. In: CG International '87 on Computer Graphics 1987. Springer-Verlag New York, Inc., Karuizawa, Japan, New York, NY, USA, pp. 25–44 (1987). http://dl.acm.org/citation.cfm?id=30300.30303
20. Kaiser, K., Miksch, S.: Information extraction – a survey. Technical report Asgaard-TR-2005-6, Vienna University of Technology (2005)
21. Kandlhofer, M.: Einbindung neuer Visualisierungskomponenten in ein Multiple Coordinated Views Framework, Endbericht Master-Praktikum (2008)
22. Kapler, T., Wright, W.: Geo time information visualization. Inf. Vis. **4**, 136–146 (2005)
23. Keim, D.A., Mansmann, F., Oelke, D., Ziegler, H.: Visual analytics: combining automated discovery with interactive visualizations. In: Discovery Science, LNAI, Springer Berlin/Heidelberg, Budapest, Hungary, pp. 2–14 (2008)
24. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: scope and challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining, pp. 76–90. Springer, Berlin/Heidelberg (2008)
25. Kienreich, W., Seifert, C.: An application of edge bundling techniques to the visualization of media analysis results. In: Proceedings of the International Conference on Information Visualization, London. IEEE Computer Society Press (2010)

26. Kienreich, W., Zechner, M., Sabol, V.: Comprehensive astronomical visualization for a multimedia encyclopedia. In: International Symposium of Knowledge and Argument Visualization; Proceedings of the International Conference Information Visualisation, Zurich, pp. 363–368. IEEE Computer Society (2007)
27. Krishnan, M., Bohn, S., Cowley, W., Crow, V., Nieplocha, J.: Scalable visual analytics of massive textual datasets. In: IEEE International Parallel and Distributed Processing Symposium, 2007. IPDPS 2007, Long Beach, pp. 1–10 (2007)
28. Lex, E., Seifert, C., Kienreich, W., Granitzer, M.: A generic framework for visualizing the news article domain and its application to real-world data. J. Digit. Inf. Manag. **6**, 434–441 (2008)
29. Muhr, M., Kern, R., Granitzer, M.: Analysis of structural relationships for hierarchical cluster labeling. In: Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), SIGIR '10, Geneva, pp. 178–185. ACM, New York (2010)
30. Muhr, M., Sabol, V., Granitzer, M.: Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets. In: IEEE International Workshop on Text-Based Information Retrieval; Proceedings of the International Conference on Database and Expert Systems Applications, Bilbao (2010)
31. Müller, F.: Granularity based multiple coordinated views to improve the information seeking process. Ph.D. thesis, University of Konstanz, Germany (2005)
32. Muthukrishnan, P., Radev, D., Mei, Q.: Edge weight regularization over multiple graphs for similarity learning. In: IEEE 10th International Conference on Data Mining (ICDM), 2010, Sydney, pp. 374–383 (2010). doi:10.1109/ICDM.2010.156
33. Rennison, E.: Galaxy of news: an approach to visualizing and understanding expansive news landscapes. In: Proceedings of the ACM Symposium on User Interface Software and Technology, UIST '94, Marina del Rey, pp. 3–12. ACM, New York (1994)
34. Ribeiro-Neto, B., Baeza-Yates, R.: Modern Information Retrieval: The Concepts and Technology Behind Search, 2nd edn. Pearson Education, Ltd., Harlow, England, Addison-Wesley (2011). http://dblp.uni-trier.de
35. Risch, J.S., Rex, D.B., Dowson, S.T., Walters, T.B., May, R.A., Moon, B.D.: The STARLIGHT information visualization system. Readings in Information Visualization, pp. 551–560. Morgan Kaufmann, San Francisco (1999)
36. Saaty, T.L.: Principia Mathematica Decernendi: Mathematical Principles of Decision Making, 1st edn. RWS Publications, Pittsburgh, PA, USA (2010)
37. Sabol, V., Kienreich, W., Muhr, M., Klieber, W., Granitzer, M.: Visual knowledge discovery in dynamic enterprise text repositories. In: Proceedings of the International Conference Information Visualisation (IV), pp. 361–368. IEEE Computer Society, Washington, DC (2009)
38. Sabol, V., Syed, K., Scharl, A., Muhr, M., Hubmann-Haidvogel, A.: Incremental computation of information landscapes for dynamic web interfaces. In: Proceedings of the Brazilian Symposium on Human Factors in Computer Systems, Barcelona, Belo Horizonte, Brazil pp. 205–208 (2010). http://dblp.uni-trier.de/db/conf/ihc/ihc2010.html#SabolSSMH10
39. Scharl, A., Tochtermann, K.: The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society (Advanced Information and Knowledge Processing). Springer, New York/Secaucus (2007)
40. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
41. Seifert, C., Granitzer, M.: User-based active learning. In: Fan, W., Hsu, W., Webb, G.I., Liu, B., Zhang, C., Gunopulos, D., Wu, X. (eds.) Proceedings of the International Conference on Data Mining Workshops (ICDM), Sydney, pp. 418–425 (2010)
42. Seifert, C., Lex, E.: A novel visualization approach for data-mining-related classification. In: Proceedings if the International Conference on Information Visualisation (IV), Barcelona, pp. 490–495. Wiley (2009)
43. Seifert, C., Lex, E.: A visualization to investigate and give feedback to classifiers. In: Proceedings of the European Conference on Visualization (EuroVis), Berlin (2009). Poster

44. Seifert, C., Kump, B., Kienreich, W., Granitzer, G., Granitzer, M.: On the beauty and usability of tag clouds. In: Proceedings of the International Conference on Information Visualisation (IV), London, pp. 17–25. IEEE Computer Society, Los Alamitos (2008)
45. Seifert, C., Sabol, V., Granitzer, M.: Classifier hypothesis generation using visual analysis methods. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) Networked Digital Technologies. Communications in Computer and Information Science, vol. 87, pp. 98–111. Springer, Berlin/Heidelberg (2010)
46. Seifert, C., Kienreich, W., Granitzer, M.: Visualizing text classification models with Voronoi word clouds. In: Proceedings of the International Conference Information Visualisation (IV), London (2011). Poster
47. Shalev-Shwartz, S., Singer, Y., Ng, A.Y.: Online and batch learning of pseudo-metrics. In: International Conference on Machine learning (ICML), Banff, p. 94 (2004)
48. Shneiderman, B.: Inventing discovery tools: combining information visualization with data mining. Inf. Vis. **1**(1), 5–12 (2002)
49. Shneiderman, B., Plaisant, C.: Designing the User Interface: Strategies for Effective Human-Computer Interaction, 5th edn. Addison-Wesley Publ. Co., Reading, MA, p. 606 (2010)
50. Thomas, J.J., Cook, K.A. (eds.): Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Computer Society, Los Alamitos (2005)
51. Tochtermann, K., Sabol, V., Kienreich, W., Granitzer, M., Becker, J.: Enhancing environmental search engines with information landscapes. In: International Symposium on Environmental Software Systems, Semmering. http://www.isess.org/ (2003)
52. Tukey, J.W.: Exploratory Data Analysis, 1st edn. Addison Wesley, Massachusetts (1977)
53. van Ham, F., Wattenberg, M., Viegas, F.B.: Mapping text with phrase nets. IEEE Trans. Vis. Comput. Graph. **15**, 1169–1176 (2009)
54. Weber, M., Alexa, M., Muller, W.: Visualizing time-series on spirals. In: IEEE Symposium on Information Visualization, 2001. INFOVIS 2001, San Diego, pp. 7–13 (2001)
55. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)