

Visual Search and Analysis in Complex Information Spaces – Approaches and Research Challenges

D. W. Fellner, T. Schreck, T. von Landesberger and J. Kohlhammer

Abstract One of the central motivations for visual analytics research is the so-called information overload – implying the challenge for human users in understanding and making decisions in presence of too much information [37]. Visual-interactive systems, integrated with automatic data analysis techniques, can help in making use of such large data sets [35]. Visual Analytics solutions not only need to cope with data volumes that are large on the nominal scale, but also with data that show high *complexity*. Important characteristics of complex data are that the data items are difficult to compare in a meaningful way based on the raw data. Also, the data items may be composed of different base data types, giving rise to multiple analytical perspectives. Example data types include research data compound of several base data types, multimedia data composed of different media modalities, etc.

In this paper, we discuss the role of data complexity for visual analysis and search, and identify implications for designing respective visual analytics applications. We first introduce a data complexity model, and present current example visual analysis approaches based on it, for a selected number of complex data types. We also outline important research challenges for visual search and analysis we deem important.

Dieter W. Fellner
Technische Universität Darmstadt, Germany. e-mail: d.fellner@gris.tu-darmstadt.de

Tobias Schreck
Universität Konstanz, Germany. e-mail: tobias.schreck@uni-konstanz.de

Tatiana von Landesberger
Technische Universität Darmstadt, Germany. e-mail: ttekusov@gris.tu-darmstadt.de

Jörn Kohlhammer
Fraunhofer IGD, Darmstadt, Germany. e-mail: joern.kohlhammer@igd.fraunhofer.de

1 Introduction

Visual-interactive techniques, combined with intelligent data analysis methods, can be valuable tools for obtaining useful insights and actionable findings from *large* and *complex* data sources [35]. They are used in many application areas including biology, medicine, engineering and manufacturing, finance, and scientific research, just to name a few. While data size relates to the nominal quantity of data under concern (such as the number of objects), complexity is related to inherent properties of the data.

The **need for visual analysis of large and complex data** stems from the general assumption that the analysis of larger and more complex data may lead to more insight (i.e., discoveries of new previously unknown knowledge). This assumption holds if the tools for analyzing these data enable the user to discover all included insights. However, creating tools scaling up with data size and data complexity is still a key challenge in the Visual Analytics area [35, 21].

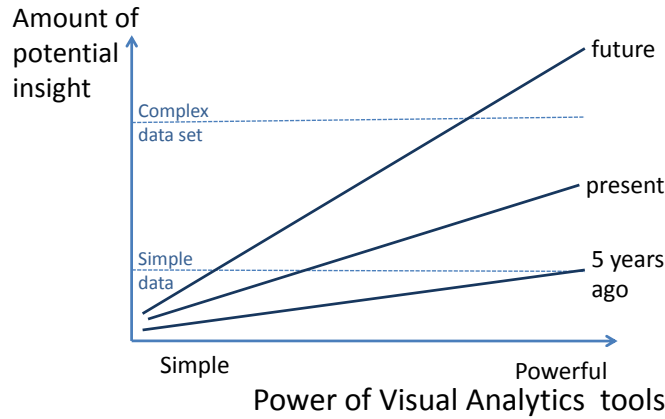


Fig. 1 Supposed functional dependency between sophistication of Visual Analytics tools and potential insight achievable, for data of different complexity. Research in Visual Analytics for complex data aims at increasing the slope of the assumed functional dependency.

While no generally acknowledged **definition for complexity** exists, we associate with it (a) the data items being difficult to compare based on raw data, and / or (b) data compound of several base data types. An example of complex data difficult to compare is multimedia data. Two raster (pixel) images typically cannot be meaningfully compared based only on the raster representations, but rather, content-based descriptions need to be extracted beforehand for this purpose. An example for a compound-complex data set is earth-observation research data, which may comprise remote sensing image data, annotated by textual meta data, and connected to time series of environmental observation parameters such as temperatures, radiation levels, humidity, or the like.

Complexity properties affect the data processing throughout the whole **analytical workflow**. Both the Visual Analytics reference model [23] and the Information Visualization reference model [14] suggest to transform input data for mapping them to visual representations. For complex data, this transformation is often difficult and ambiguous. Usually, domain- and data-specific transformation steps need to be applied, to make the data available for visual mapping and aggregation. Moreover, user interaction methods, visual displays and further automatic data analysis methods need to be adapted to the complexity characteristics of data.

Given several different notions of data complexity and their implications for the Visual Analysis workflow, there is a need to more explicitly consider the role of data complexity in Visual Analytics. We here examine two important key **user tasks** in Visual Analysis systems for complex data: *Searching* for data items of interest, and *analyzing* for relationships among and between sets of data items. Searching and analyzing are very related, and often, a sequence of searching tasks is conducted that leads to findings on the global analysis level.

Progress in data acquisition, storage, and transmission leads to data repositories integrating different data modalities. To date, many visual analysis systems focus on data sets of given complexity, mostly addressing a single complex data type. The amount of potential insight obtained from data of a given complexity can be raised by the degree of sophistication of the visual analysis system. A visual analysis tool of given technological development status will provide increasing insight potential, as data gets more complex. However, the relation between sophistication of the visual analysis solutions and the complexity of the data is limited by the technological state of the art of the tools. We expect that by systematically researching and improving visual analysis technology, the slope of the relationships between tool sophistication and data complexity with respect to potential insight can be increased. Figure 1 sketches the basic functional dependencies as we assume them. By pushing the limit in visual analysis tool support for complex data, the slope of the functional relationships between the two variables can be made more steep. Focusing on improved approaches for addressing data complexity in visual analysis tools, we hope to be able to push the limits. In the following, we give a definition of complex data sets, provide an overview of approaches to visual search and analysis therein, and identify future research challenges. We state that if the raised challenges are addressed appropriately, future visual analysis tools will be able to derive more potential insight from a given type of complex data.

The remainder of this paper is structured as follows. In Section 2, we identify two main sources for complexity, and discuss their role in relationship to Visual Analytics applications. In Section 3, we discuss tasks and problems when dealing with complex data. Section 4 then provides proposed Visual Analytics solutions to a number of example problems from various data and application domains. Based on this, in Section 5, we identify a number research challenges considered important and interesting. Finally, Section 6 concludes.

2 Definition of Complex Data Sets

The term complex data is often used in a loosely defined way. In our view, data complexity can be attributed to two fundamental dimensions. Complexity may stem (a) from intrinsic properties of a given single kind of data (type complexity), or (b) from the data being structured based on a mix of different types, either simple or complex in turn (compound complexity) (see Figure 2). Both dimensions impact the difficulty of algorithmic and visual processing of the data. For searching and analyzing, a key fundamental data operation involves calculation of similarity properties among data items.

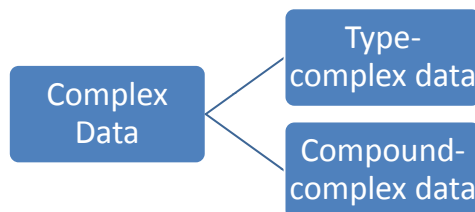


Fig. 2 Categorization of complex data. In our definition, we distinguish between complexity stemming from inherit properties of a given single data type (type-complex data), or from the data being compound of multiple base data types (compound-complex data).

In the first type of complex data (**type-complex data**) the specific properties of a single given data type incurs difficulty to algorithmically process the data for similarity in a straightforward way. In particular, for these data types it is difficult to measure similarity meaningfully between data instances based on the raw data representation. For example, image data, audio data, video data, biochemical molecule data, 3D object data, or graph data are data types of this category. In all cases, the raw data needs to be transformed by specialized preprocessing steps for further algorithmic and visual analysis. The difficulty can be explained in the following example showing the difference between multivariate numeric data (considered simple here) and image data (considered complex here). For multivariate data, we can usually compute the similarity of data records based on forming sums of absolute differences of the respective field values in the records. In contrast, consider the task of comparing a query and a candidate image. For most practical purposes, it is not possible to calculate the similarity based on the raw image pixel arrays. Rather, a preprocessing step which extracts relevant information from the images, such as the presence of specific object types, or color and texture patterns, is needed. Then, query processing can take place on this extracted information [30].

In the second type of complex data (**compound-complex data**) the data items are aggregated from multiple “base” data types, each of which could be complex or non-complex in turn. Again here, it is difficult to calculate similarities, because it is a priori not clear how to aggregate the similarities between the individual data

components. Depending on the application context, either one of the base data types could be the relevant perspective, or also, a combination thereof. Moreover, the data components may be complex themselves, raising the complexity by the number of involved base data types. As an example, consider research data from the earth observation domain. Here, measurement data can be compound of several complex data types. A realistic example includes multivariate time-dependent measurements of environmental parameters. Additionally, geo-locations, trajectories, and image data may be available for the specific data. Even further, the particular observation experiment which lead to the acquisition of this data may be described in a research paper, which is relevant to understanding the data. An earth observation scientist might be interested to search and analyze for each of these aspects simultaneously. Nowadays, large repositories of such data are set up and made publicly available. While access to the data per se is given, in absence of appropriate search and analysis tools, often these repositories are not accessible to the best possible extent. Earth observation data is just one example for compound-complex data. Others include compound graphs, biologic experimental data, spatio-temporal data, intelligence data compound by textual reports, intelligence findings and image documentation, and so on and so forth.

3 Tasks and Problems of Visual Search and Analysis in Complex Data

We next describe two fundamental user tasks in visual analysis systems - searching and analyzing. We then outline the key problems of supporting these tasks in presence of complex data.

3.1 *Visual Search and Analysis*

Searching and analyzing are key user tasks in information systems. Searching relates to finding information entities of interest to a user on a more local level, based on specific query formulation. Analyzing, in its generic sense, can be defined as finding structures and abstractions on the set level, adding to the understanding of a data set as a whole.

Search is an inherent part of the data analysis process and can take several **forms**. Although it may not be seen at the first sight, search tasks are comparable to the basic information visualization tasks defined by Shneiderman [32]. Search includes, e.g., identification of the data needed for the analysis, searching for similar data items among a set of items, detection for recurring motifs in a sequence or network, or discovery of outliers or exceptional events. For visual support of these tasks, appropriate user interfaces are needed. These interfaces need to include e.g., visual means of query specification and results presentation.

By *analysis* we understand tasks related to identification of global relationships and structures in the data. Questions of interest relate to the number of groups existing in the data, the similarities and differences between them, and how they relate to each other. Cluster analysis and association rule mining are two examples of analysis methods. Visual support for analysis tasks require the appropriate visual steering of the analysis algorithms, and expressive visual displays to present the output of the analysis methods. Also, navigation facilities to allow overview and details-on-demand are important ingredients in respective systems.

Searching and analyzing are often interrelated, with cycles of corresponding activities. For example, evidence for a hypothesis may be collected by issuing a series of queries, which select subsets of the data for some aggregation purpose. Often, sets of searching subtasks, each of relative short duration, are nested within longer-running, overarching analysis processes.

3.2 Problems in Presence of Complex Data

Complex data imply specific problems for visual search and analysis. These problems depend both on the task (searching vs. analyzing) and on the type of data complexity. For searching, Figure 3 illustrates a process model of visual search in complex data. Based on visual query specification, and by means of appropriately selected and configured descriptors, similar objects to the query are identified. Visual overview displays are useful for showing search results in context of the whole data set. In our model of the analysis process (see Figure 4), first the suitable data set needs to be found for further processing. After determining the appropriate similarity function and selecting an analysis method, the results are visually inspected. This process includes several feedback loops creating an incremental process. In both cases, each process step poses problems for developing visual analysis methods.

For *type-complex data*, sophisticated data transformation needs to be applied before data items can be queried, compared, grouped, and visualized. Typically, custom similarity functions, for example, based on descriptors (or feature vectors) need to be defined. However, for many type-complex data sets, multiple alternating descriptors are possible, and it typically is not clear which one suites the current task best. Furthermore, descriptor extraction is a non-trivial, parameterized process, and analysts are often not aware of its implications. On the other hand, meaningful analysis (interpretable results) requires the user to understand the specific notions of similarity which lead to search results or object groupings. Also, query specification is difficult if descriptors cannot be easily interpreted - direct numeric feature specification is typically not useful for average users. How can querying be visually supported, visually representing the relevant similarity concepts and thresholds for which objects are considered similar by the system? How sensitive are query and analysis results with respect to parameterizations of the similarity function? Such questions need to be addressed by ideal visual analysis systems.

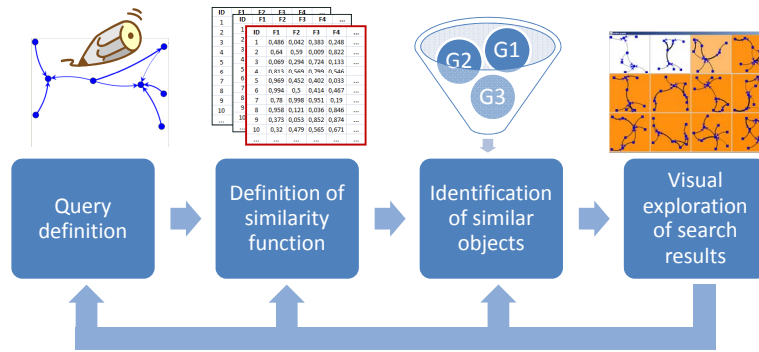


Fig. 3 Visual search process. By means of a visually specified query, and relying on a selected descriptor, similar objects are retrieved and visualized, often in context of the whole data repository.

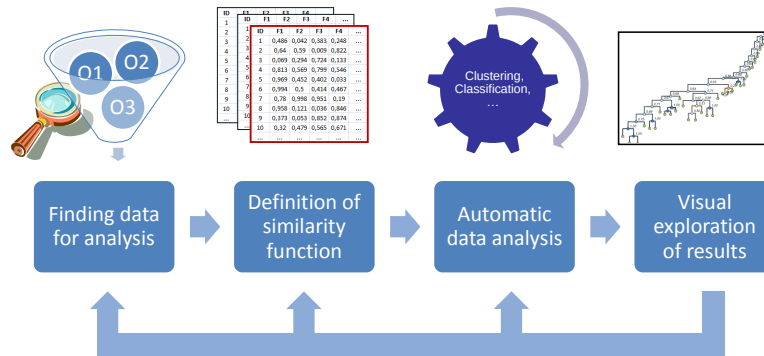


Fig. 4 Visual analysis process. After specification of the data for the analysis and their similarity function, the selected analysis function is executed and the results are visually inspected.

These problems are potentially multiplied for *compound-complex data*. In these data, searching and analyzing tasks need to be based on a joint similarity function which appropriately reflects similarity concepts for each included base data type. All base types have possibly different similarity functions. These similarity notions need to be both configured individually and appropriately aggregated. Consider the example of compound-complex research data from our example of earth observation research, where an exemplary task is to analyze for similar observations. Similarity can be constituted by similarity of respective measurement series, but also, geographic location, measurement method applied, or researchers involved. How can an aggregate similarity function be defined for searching in such data? How can such data be clustered? Clearly, the user needs to be given appropriate visual query formulation tools which allow to select and weight the involved data perspectives, and specify a query in all of these relevant perspectives.

We summarize key design problems for visual search and analysis systems for complex data as follows. As can be seen from the respective processes, there are many parallels between the two tasks (search and analysis) but they also have several specifics.

Visual Search

- *Query formulation.* The user needs to be enabled to specify the query visually and interactively. The design problem is to derive visual representations for the query properties of interest. In case the data has a visual representation, so-called *query by sketch* is possible, where the user outlines a draft shape to be searched for. A problem to address is the level of abstraction, by which the query is specified. In case of compound-complex data, the query specification is potentially multiplied by each involved data modality.
- *Similarity function.* The similarity function to use for evaluating the query needs to be selected and parameterized. For the user to make an informed selection, the system should visually represent the implications of the selected similarity function for the result to be expected. While for type-complex data a single similarity function needs to be specified, for compound-complex data, possibly for each base data type one similarity function needs to be selected, and a combination needs to be found.
- *Visual result presentation.* The visual search systems needs to present the sequence of found results and their potential relevance to the query. Each object needs to be shown by a visual representation. In case of data with visual representation, thumbnail views are common. Visualizing result sets for compound-complex data involves finding appropriate visual representations for the combined data perspectives, and how each of the base data instance for each data type relates to the issued compound query.
- *User feedback.* Effective search systems require the user to quickly converge to a satisfying result, only using a few iterations of query adaptation and result inspection. Therefore, it is crucial that the system offers ways for the user to understand why the found results relate to the user query, both in terms of the query specification and descriptor selected.

Visual Analysis

- *Similarity function.* Like in searching, many analysis algorithms rely on a similarity function to be defined for the data objects (e.g., clustering). This should be also supported by involving the user in an interactive process of defining similarity and evaluating analysis results. It again often involves selection of an appropriate descriptor, and specification of combinations of descriptors, in case of compound-complex data.

- *Selection of analysis method.* The user needs to interactively select an appropriate analysis algorithm to apply. This involves selecting the type of analysis algorithm (e.g., cluster analysis, association rule mining, classification analysis, etc.) as well as its configuration. This is not an easy task as not all analysis methods can deal with complex data sets and therefore specific analysis methods need to be applied or developed.
- *Visual result presentation.* Presentation of analysis results, similar to presentation of search results, requires finding an appropriate visual abstraction. While in search, the level of interest is on the object level, on the other hand in analysis, often aggregates (e.g., clusters) or abstractions (e.g., hierarchies) are found. These need to be visualized, reflecting possible visual representations of the single of compound base data types.
- *User feedback.* In analysis algorithms, user feedback again plays an important role. We expect that to arrive at satisfying results, several analysis iterations need to be performed. Comparison of search results is rather straightforward, as ranked lists need to be compared. In case of analysis, the problem may become more difficult, as aggregate and abstract analysis outputs need to be compared, for the user to understand the differences between the choices. For example, an appropriate visualization should allow the user to effectively compare two clusterings obtained from two different compound similarity functions.

4 Approaches

In this section, we discuss selected examples of visual search and analysis systems, which illustrate the variability of the problem. In Section 4.1, we will illustrate key principles by means of classic example systems from the field. In Section 4.2, we will discuss some approaches for type-complex data, and in Section 4.3, we will present examples for support of compound-complex data.

4.1 Generic Examples for Visual Search and Analysis Systems

Generic approaches to visual search and analysis date back as early as to the beginning of Information Visualization as a field. Shneiderman in his Visual Information Seeking Mantra [32] proposed to support the search and analysis process by visual-interactive tools. He and Ahlberg proposed the FilmFinder system [3], which supported a new way of interactive search (see Figure 5(a)). In this concept, visual overviews allow to analyze the data set at an abstract level, with interactive query interfaces allowing drill-down queries to arrive at details-on-demand. Visual-interactive displays for searching and analyzing aim to provide intuitive access and navigation. Leveraging the human visual perceptual system, they are supposed to

provide a high bandwidth interface, encourage explorative analysis and creative processes in the users mind.

Another generic example of an exploratory system for complex data is the well-known INSPIRE system for exploration of document collections proposed by Thomas et al. [36] (see Figure 5(b)). Text is type-complex data as it cannot be meaningfully compared based on the raw data, but it needs to be preprocessed e.g., using word frequency vectors. INSPIRE relies on projection of high-dimensional document vectors to a 2D display, to overview document corpora for similarity of topics and for exploration. An appropriate visual design shows documents in a landscape metaphor which can be readily navigated by the user.

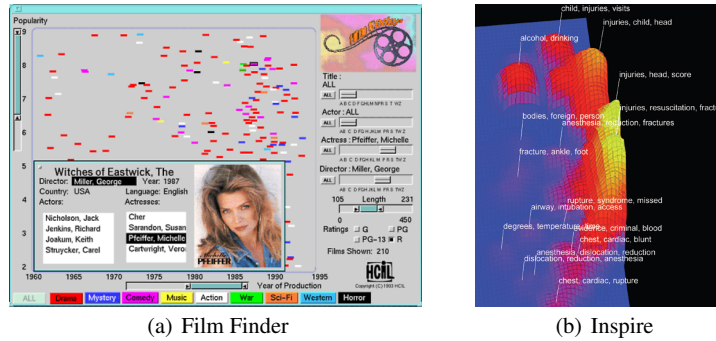


Fig. 5 Left: The Filmfinder system [4] is an example for visual search in multivariate data (Figure ©1994 ACM, Inc. Included here by permission.). Right: The Inspire system [36] is an example of an analysis system, which allows to relate and compare subsets of elements in a visual way (Figure ©1995 IEEE).

We can learn from examples such as these, that complexity is often dealt with by **simplification**: Complex data is transformed to feature vectors; dimensionality reduction is applied to project data to interactive displays, and large data sets are sampled to provide overviews. In case of compound-data, projection to a selected data perspective of interest is a pragmatic approach. However, such approaches often incur a loss in formation. Visualization also applies a simplification, in mapping only selected dimensions to visual variables, or visually aggregating many data samples (such as many documents in INSPIRE) to a landscape, where height indicates groups of data and position their relationships.

There are general principles that help the user working with complex data. These include **consistency** and **user guidance**. Consistency means the same way of working with similar data in various environments. This is especially needed when dealing with compound-complex data in multiple perspectives. There, each perspective should use the same interaction means and mappings, if possible. User guidance helps the user in the search and analysis process providing her with a set of steps to follow, or recommendations for suitable parameters in algorithmic analysis.

4.2 Example Approaches to Visual Search and Analysis of Type-Complex Data

In the next section, we consider examples for visual search and analysis in type-complex data including 3D object data, graph data, and biochemical data.

4.2.1 Visual Search in 3D Object Data

Many multimedia data types are of type complexity or compound complexity. An example of type-complexity is the area of 3D model data. By specific data structures, the shape and other properties of 3D objects can be modeled with applications in computer-aided manufacturing, architecture, and simulation, just to name a few. A widely used data structure to encode the shape of 3D models are polygonal meshes. While simple in terms of data structure, two mesh models cannot be meaningfully compared based on their polygons. However, a wealth of description extraction methods has been proposed to date [34]. The idea is to extract descriptor from mesh models, which allow for meaningful comparison. In the project PROBADO, we have considered visual search methods to help architects query in 3D building models. The idea is to allow the users to quickly specify properties of interest in a building. To this end, we considered two query modalities: Querying by global 3D shape, and querying for room connectivity structure [11]. Querying by global building shape is supported by the user sketching the coarse outline of a building by a simple 3D block editor. Having entered a 3D sketch, any global 3D descriptor [34] can be used to retrieve similar objects. A more specialized query modality suggested by architects included the querying for the structure of rooms in a building (floor plan). To this end, we devised a method to extract a room connectivity graph from each building in the repository. The user then enters a query structure by means of a simple graph editor, and again, the system finds similar objects based on a graph matching strategy. Figure 6 illustrates a sample 3D building model, and the two query editors.

From this example, we see that often, many generic descriptors already exist for a given complex data type. However, not always do the existing descriptors support all possible domain-specific search modalities. For example, the room connectivity structure was of interest to architect users, so it needed to be developed anew.

4.2.2 Visual Search in Graphs – Visual Query Definition

Graphs are used in various application areas such as chemical, social or shareholder network analysis. Finding relevant graphs in large graph databases is thereby an important problem. Such search starts with the definition of the query object. Defining the query graph quickly and effectively so that it matches meaningful data in the database is difficult. In [27], we introduced a system, which guides the user through

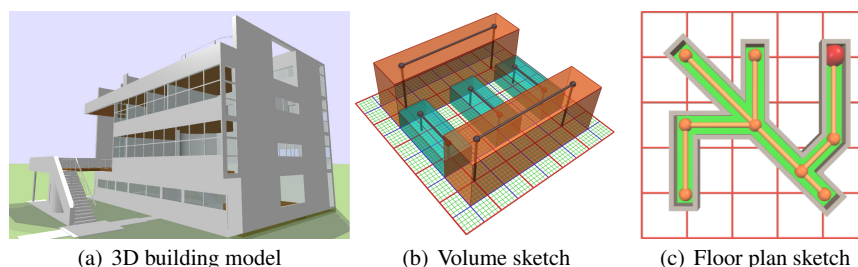


Fig. 6 Example visual search modalities for querying in 3D architectural object data.

the process of query graph building. We proposed three ways of defining the query graph, which support the user with intelligent, data dependent recommendations. In this way, the query graph is defined more quickly and corresponds better to the underlying data set.

1. Smart Choice of Data Samples

The first approach employs a query-by-example technique, where one existing graph is used as a query object. For the choice of query object, we offer a suitable selection of example graphs from the database. The proposed selection provides an overview of the available graphs. It is based on the result of clustering by Self Organizing Map (SOM) algorithm [24] as introduced in [26] (see Figure 7 left top).

2. *Graph Sketching Supported by Data-Dependent Graph Building Blocks* Another approach to query definition is query-by-sketch – creating the query object itself. Graph editing from scratch by adding individual nodes and edges one by one can be very time consuming for large graphs. Therefore, we extended graph editing by adding multiple nodes and edges at once – using the so called graph building blocks. The building blocks are small sub-graphs that occur often in graphs (i.e., motifs). These blocks are interactively combined so they support fast creation of graphs. Moreover, we analyze the underlying data space to present to the user additional guidance, in particular, information on frequency of occurrence of these blocks in the database (see Figure 7 left bottom).

3. *Combination of Sketching and Examples* As sketching may be time consuming and examples may not provide enough flexibility, we combined both approaches. The query definition starts from an existing object chosen from the proposed set. This object can be modified by adding and deleting of edges and nodes or adding building blocks. Combination of these techniques provides a fast definition of a specific and meaningful query object (see Figure 7).

4.2.3 Visual Search and Analysis of Biochemical Data

– Similarity Function Definition using Visual Comparison of Descriptors

The analysis of biologic and chemical data gains on importance in the Visual Analytics community. Biologic and chemical data can be regarded type-complex data

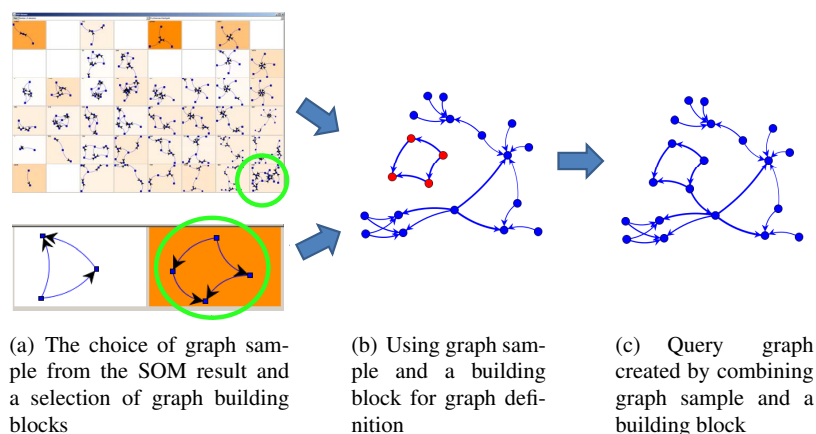


Fig. 7 Combining smart sketching with data samples for leveraging the advantages of both techniques. a) The proposal for graph samples using SOM clustering and graph building blocks with frequency indication. b) The selection of a graph sample and a building block for creating the query graph. The selected sample and the building block are highlighted with green circles in the proposal view. c) The final query graph combining both graph samples and building blocks with sketched edges.

types. For example, chemical compounds cannot be analyzed directly but need to be described by their properties such as size, charge, solubility, atom connectivity, etc. The selection of these properties is used for defining similarity between objects. The employed description is heavily use case dependent, therefore, user knowledge is very important for the evaluation of the selections. This evaluation is however difficult, if different representations of the whole dataset are presented to the user. For example, in the analysis of high throughput screening (HTS), an overview and comparison of thousands of molecules is needed.

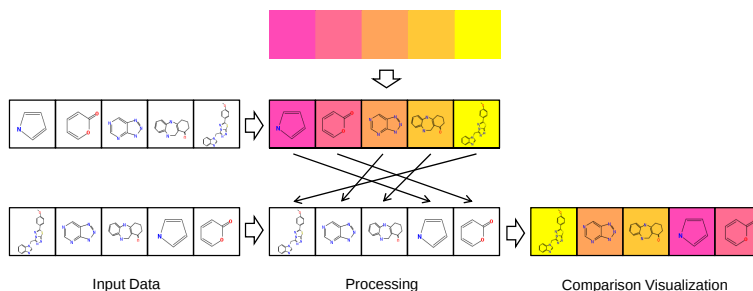


Fig. 8 Two meaningful data descriptors of biochemical data and their comparison. Descriptors: atom resp. nitrogen count. Left: The input data is sorted according to each descriptor. Center: Color is mapped to the first ordering. The sorting is compared using connectors. Right: Compact comparison view based on object identity reveals descriptor correspondence.

In [13], we presented a novel visual analysis approach for determining data descriptions suitable for the task at hand. We developed dedicated visualizations for comparison of sets of multi-dimensional data descriptors. These techniques are based on low-dimensional data presentation using color for comparison of groupings resulting in the different descriptor spaces (see Fig. 8). For large data sets, we employ adaptive grids with clustering properties (Self-Organizing Maps, [24]). These views allow for spotting overall similar descriptors and locally similar object groups in heterogeneous data sets. The finding of potentially interesting descriptors is supported by an interactive pipeline, which guides the user through the analysis process. The result of initial automatic data analysis provides recommendations and offers the user the possibility to interactively refine the results. These refinements are supported in visual-interactive way.

As an application example, 18 commonly used chemical descriptors for 9989 molecules with 773 dimensions in total were examined. The comparison of two descriptors in Fig. 9 shows a very homogeneous color gradient representing the descriptors for weight and number of atoms of the molecules. This validates an expectation of the coherence between weight and size. Looking at the comparison of the ExtendendFingerprint with the WienerNumber descriptor, we see that many cells are homogeneously colored (Fig. 9 right). All of the purple molecules in the WienerNumber SOM are located in one cell of the ExtendendFingerprint SOM. If the pharmacologist is interested in these molecules, the WienerNumber descriptor is preferable. It leads to a higher diversity of the concerned molecules at a lower dimensionality (1 vs. 26).

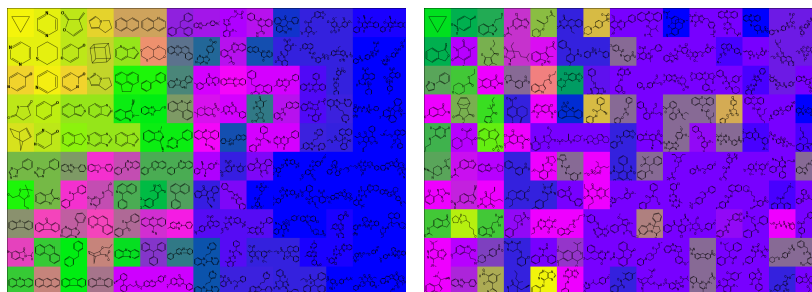


Fig. 9 Left: Comparison of the weight to an atom count descriptor. The homogeneous color gradient validates the expected correlation of the descriptors. Right: The 1-D WienerNumber descriptor shows a high separability for molecules which are all in one cell in the SOM of the 26-D ExtendendFingerprint.

4.3 Example Approaches to Visual Search and Analysis of Compound-Complex Data

We next discuss example systems for search and analysis in compound-complex data. Examples span research data, geo-temporal event data, and security-related data.

4.3.1 Visual Search in Research Data

– Visual Query Definition and Visualization of Search Results

Science as a domain heavily depends on the timely availability of appropriate information. Recently, the need for persistent storage of data produced in public research has been recognized by data producers, researchers, and public funding agencies alike. For example, in the earth observation sciences, massive amounts of data are collected by sensor networks, or by data acquisition campaigns. Currently, large data repositories, such as the *PANGAEA Publishing Network for Geoscientific and Environmental Data* [28], are being built. Persistent availability and sharing of such data among the research community can foster scientific progress, allow to reproduce results, and document present states of the environment for future research.

Visual search and analysis facilities in such data are highly desirable to enable researchers to access the data [5]. The data under concern typically is complex in that it consists for multiple base data types which together form the observation data. For example, earth observation data can consist of time-dependent multi parameter measurements for environment factors, in addition to images of multi spectral satellite analysis and X-ray images of sediment cores, extracted from the area of measurement. Typically, researchers want to search for content of the data, to compare or formulate hypotheses. To this end, the research data needs to be indexed by an array of different descriptors; and appropriate visual search interfaces need to be provided. In [10], we have described an early prototype system which allows a multi-faceted search in earth observation data. Content-based search is supported by allowing the user to specify the draft shape of a time series of a given observation parameter (cf. Figure 10 (left)). Based on curve descriptors, the most similar curves can be retrieved, and further filtering of result sets based on geo-location, seasonal and other meta data attributes are possible. The search results and refinements thereof are visualized in context of an overview of a larger data set, e.g., the given data repository (cf. Figure 10 (right)). To this end, a visual cluster analysis of the overall data set is performed, and search results are highlighted in their context.

In this system, we have explored the tight integration of searching and analyzing. A visual catalogue is the central visual element of the system, showing an overview over the most important time series patterns. Search results can be shown in context of the overview. Also, curves from the overview can be selected and adapted in the query editor, for an adjusted search. While the system also allows to query for

the other involved compound data aspects (textual meta data, geo-location, etc.) currently the system is oriented towards search in the time series shape space.

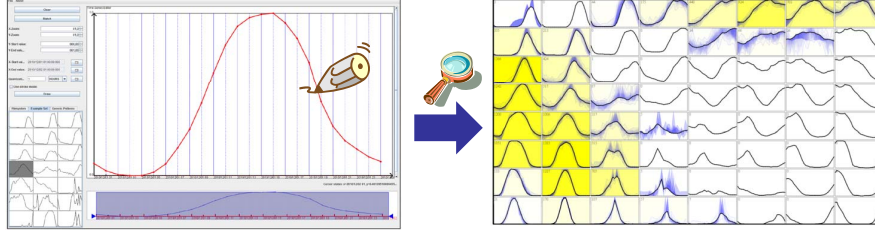


Fig. 10 Visual query specification (left) and result visualization (right) for searching in time-oriented research data.

4.3.2 Visual Search and Analysis of Spatio-temporal data – Identification of Interesting Events

The analysis of spatio-temporal data plays a prominent role in many applications such as transportation, meteorology, finance or biology. One area is the analysis of movement (i.e., trajectory) data. For example, car movement for traffic monitoring, animal behavior in biologic observations, people movement in emergency situations, or dynamics of stocks on the stock market for financial investment decisions. Movement data is a compound-complex data type composed of two data types: time and location. The analysis of these trajectory data is a well studied problem in the visual analytics area [9, 29, 20, 7, 8, 15]. Movement data can be studied for individuals or for groups of individuals. In [12] we propose an approach that addresses the analysis of grouped spatio-temporal data. It is based on the notion of Parallel Sets [25], extended for automatic identification of interesting points in time that are suggested to the user for inspection.

Generally, the groupings data may be pre-defined (e.g., by identification of animal herds in biology), or may be a result of previous analysis (e.g., clustering). When the group membership changes over time, it is necessary to examine these aspects (e.g., which herds change members and when).

As the number of analyzed time moments may be very large, the group changes cannot be manually inspected in each time point. Therefore, a good selection of the points in time for a detailed analysis is important. It should represent the data well – reveal important movements or outliers. It should highlight overall trends and identify time periods of high activity (shorter intervals). Moreover, detection of outliers provides a set of moments with extraordinary group-change events.

As an application example, we can regard the analysis of people movements in the case of an emergency. As a basis for respective research, the VAST Challenge 2008 data [19] includes the movement data for 82 subjects in a building over 837

points in time. The grouping is based on areas in the building (Fig. 11 top left). It assigns subjects into groups according to their location in every time moment. In this scenario, at a specific time, a bomb detonated and afterwards people die or start to move towards the exits (turquoise and purple areas).

The automatic analysis of group changes identifies interesting time moments for detailed analysis. The result puts more emphasis on time periods of high movements (after the explosion) and identifies behavior of people who move differently from the rest or in an unexpected way (away from exits). Tracking these people reveals that despite their odd routing, the majority reaches the exits (Fig. 11 bottom).

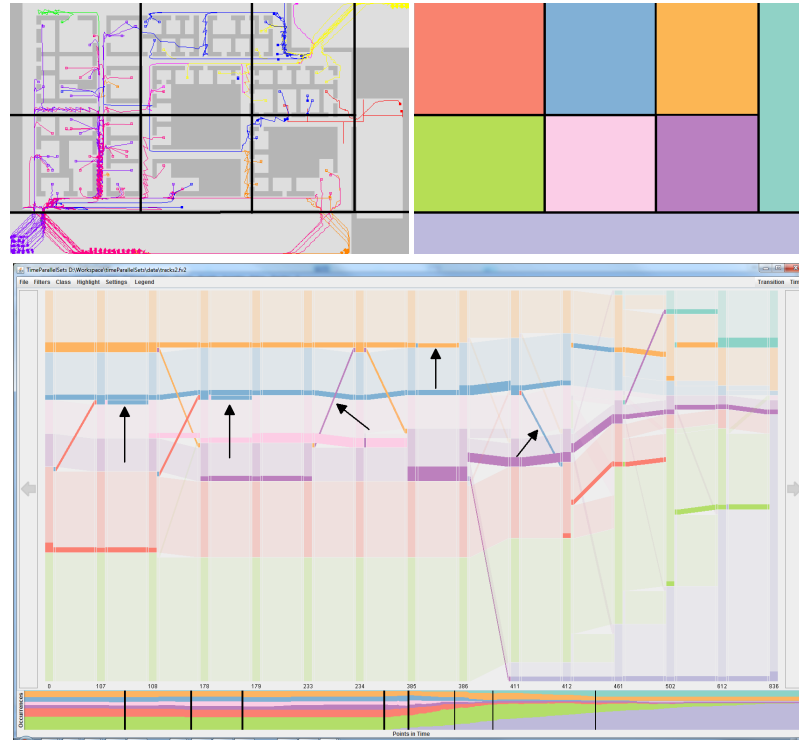


Fig. 11 Top: The example of emergency evacuation analysis using dataset from the VAST challenge 2008 [19]. Left: Trajectory based visualization of the movement of the people [6] with partitioning of the building into 8 areas. Right: Coloring of the areas. Bottom: Analysis of group changes and unexpected movements showing automatically selected time movements. Tracking of surprising movements reveals that surprisingly most of these people reach the exit in time.

4.3.3 Visual Analytics for Security

Starting from the NVAC initiative (National Visualization and Analytics Center) in the US and its visual analytics research agenda [35], a number of research programs and initiatives evolved in the direction of visual analytics for security. While in the United States there has always been an applicable emphasis on homeland security among other fields, there was no such strong focus in Europe where a large number of application areas offered a wide range of opportunities [22].

Partly initiated by Jim Thomas and after different transatlantic initiatives and joint workshops, Germany and the U.S. have jointly started the project VASA (Visual Analytics for Security Applications) to develop tools that will aid in the protection, security, and resiliency of U.S. and German critical infrastructures in 2010 [1]. The goal of VASA is to apply visual analytics to disaster prevention and crisis response, with a focus on critical infrastructures in logistics, transportation, food safety, digital networks and power grids at the national levels.

VASA works on a fundamental challenge in the analysis of compound-complex data. After all, critical infrastructures are complex socio-technical systems with components and sub-systems that are characterized by specific behaviors which result from the applied rules of physics, from technical specifications, and the established control regime. Such behaviors of single critical infrastructures are quite complex even under normal conditions, based on a variety of base data types. Interdependencies between critical infrastructures complicate the resulting behavior further through potential cascading failures and nonlinear effects. Decision making requires pre-processing of data and information which takes the specific context into account and provides the relevant information and the appropriate level of detail to the decision maker to enable effective and timely decision making. The main challenge in the VASA project is the interplay between complex information models, precise simulations, special purpose analytics, and decision making under uncertainty. All four aspects combined will enable new visual analytics systems for interdependent critical infrastructures.

Another example for visual analytics of compound-complex data is the project VIS-SENSE, funded by the EU [2]. The main goal of VIS-SENSE is the research and development of novel visual analytics technologies for the identification and prediction of very complex patterns of abnormal behavior in different application areas ranging from network information security and attack attribution to attack prediction. One important aspect of VIS-SENSE is a decision support system based on compound-complex data stemming from multiple layers of available information, ranging from low-level network data, topological network aspects, to results from network analytics. Again, similarities have to be calculated through aggregations of multiple base data types, guided by human experts - a challenge and opportunity to showcase the added value of visual analytics approaches.

5 Research Challenges

The previous examples served to illustrate the breadth and with of the problem of visual search and analysis of complex data. We believe that in accordance with increasing volumes of data, also complexity of data arises posing new challenges to the development of visual analytics tools. These challenges are strengthened by the emergence of new application areas such as biology, medicine, architecture, emergency management etc. Integrated search and analysis, in a cross-domain, cross-data-type and cross-data-repository environment will become more and more important, and will thereby require new appropriate solutions.

As an example, the definition of similarity functions for complex data can be considered. To date, already many data transformation methods are available, which allow to extract descriptors for complex data, and make them comparable in this way. Visual search and analysis systems should rely on these established methods where possible. However, it is typically difficult for a user to chose for the appropriate descriptor, either because there are too many available, or that the similarity notion required is not covered by the existing descriptors.

We next summarize a number of research challenges we deem interesting and critical, in the context of the discussed problems and application examples.

Infrastructures

Visual Analytics is a multi-disciplinary discipline, which incorporates research from various fields. Practitioners in visual analytics have started to implement ad-hoc systems, such as in-memory databases or user-steerable algorithms. However, these are still quite isolated attempts and not sustainable solutions in the long term. The community lacks an infrastructure to allow a flexible interoperability of components that might be specialized for certain type-complex or compound-complex data [21]. The goal is to allow practitioners from different fields of research to benefit more from each other's work.

This corresponds to the challenging task to design a common language, a collection of accepted practices and an architectural model that can be agreed upon by different fields of research related to data analysis. Current research in data analysis is dispersed and sometimes virtually isolated in their respective domain. In several analytics technologies, database researchers, machine-learning and data analysis researchers, as well as visualization researchers focus on specific aspects. However, visualization approaches, data management procedures, and data mining methods all have to work together in newly orchestrated ways, leading to a new definition of interoperable visual analytics building blocks that allow the coherent creation of visual analytics systems.

New Data Types

Once the building blocks of visual analytics systems are well understood, more research on data typing is needed. For example, exposing the semantic type of data in databases is essential in order to know what kind of analysis can be applied and what kind of visualization is meaningful. Today's data classifications (like nominal or quantitative) are rich enough for most statistical approaches, but it is not sufficient for visualization. The semantic web is an example of an application domain where sophisticated data types are being defined, but there are other initiatives and it is not clear yet how they will converge and how the field of visual analytics will benefit from it.

Search Problem and Comparative Visualization

Searching and analyzing in complex data require the user to make a number of profound decisions regarding query specification, descriptor selection, algorithm configuration, and combining of different data perspectives. Arriving at satisfying search and analysis results requires also to solve a meta search problem for these search and analysis parameters. Only visual systems which provide fast response times and can cope with high data complexity and sizeability allow this process to be effective. This puts high requirements with respect to scalability of the implementations.

Furthermore, appropriate visual representations are needed to show the user the implications of specific choices. How do the search result lists differ, if the descriptor is changed? How does a clustering result change with respect to the algorithm parameters set? Comparative (or delta) visualization tools could be helpful as a meta visualization.

User Guidance in the Visual Analysis Process

In the field of information and multimedia retrieval, relevance feedback [30] is a standard technique to help users indirectly to configure search parameters, e.g., choice of descriptors and similarity functions. Based on the user providing relevance votes to candidate results, an optimization problem is solved for weighing features. We believe the relevance feedback approach is a promising tool to help the user in an intuitive, indirect way to solve the descriptor and parameter choice problem. However, to be applicable to the visual search and analysis problem in complex data, we believe it needs to be adapted to reflect different data structures. Most importantly, choice of descriptors, and weighing of individual similarity functions for compound-complex data need to be optimized. To this end, the relevance feedback problem needs to be reformulated. Also, it needs to be considered what is the right level of relevance feedback judgments. Possibly, new interaction techniques need to be devised as well. While a difficult problem for visual search, relevance feedback

for visual analysis is expected to be an even harder problem. A formal model for the analysis process in complex data is needed to discuss where relevance feedback for visual analysis can be installed.

Benchmarking

Benchmarking and evaluation play an important role in devising effective visual search and analysis systems. In the area of multimedia retrieval, benchmark data sets are available mainly for standard and type-complex data such as multivariate data [18], 3D models [31] or images [16, 17]. To our knowledge, there are no established benchmarks for compound-complex data available, up to the TREC-Video data set [33]. Measuring the effectivity of visual search systems requires extended benchmark data sets, which together with user-oriented evaluation approaches are useful to compare new system designs. Benchmark data sets of general analysis problems are however expensive and difficult to obtain. The VAST analytic challenges [19] are a promising starting point to compare visual analysis systems. A deeper understanding and modeling of the analysis process could be expected to lead to more analysis benchmarks being devised in the future.

6 Conclusions

Visual search and analysis are important key tasks in making use of data. Besides nominal data volumes, data complexity is a scalability limit for existing solutions. We discussed two views on complexity in this article. One is based on the inherent complexity properties of a given data type (type-complexity), while the other stems from data being composed of several base data types (compound-complex data). Supporting visual search and analysis in this data raises several problems, including choice of data descriptors, parameterization and weighting of similarity function, visual query specification and result visualization. We aimed at illustrating the breadth and width of the problem by considering a variety of application scenarios from domains such as 3D object data, network data, scientific research data, and biochemical data. The presented solutions are just individual solutions in a large problem space. We believe that approaching a number of identified research challenges, especially in comparative visualization, user feedback, benchmarking, and infrastructure will foster further development of new solutions. Given the emergence of data in ever growing volumes and in increasing complexity, the community requires such novel approaches and solutions to access and exploit today's information spaces.

Acknowledgments

This work has been partly supported by the following research programs and projects: The project Visual Feature Space Analysis, funded by the German Research Foundation (DFG) within the Strategic Research Initiative on Scalable Visual Analytics (SPP 1335); the project VIS-SENSE funded by the European Commission's Seventh Framework Programme (FP7 2007-2013) under grant agreement nr. 257495; the THESEUS Programm funded by the German Federal Ministry of Economics and Technology; the German part of the project VASA funded by the German Federal Ministry of Education and Research; the project PROBADO funded by the German Research Foundation (DFG) within the LIS initiative; and the project VisInfo funded by the Leibniz Association (WGL). We are grateful for helpful collaboration with Prof. Kay Hamacher and other colleagues within research projects.

References

1. VASA Addresses Cascading Effects Across Critical Infrastructures. URL <http://www.theivac.org/content/vasa-addresses-cascading-effects-across-critical-infrastructures>. Last accessed on Aug 11, 2011
2. VIS-SENSE : Visual Analytic Representation of Large Datasets for Enhancing Network Security. URL <http://www.vis-sense.eu>. Last accessed on Aug 11, 2011
3. Ahlberg, C., Shneiderman, B.: Visual information seeking using the FilmFinder. In: Conference companion on Human factors in computing systems, pp. 433–434. ACM (1994)
4. Ahlberg, C., Shneiderman, B.: Visual information seeking: tight coupling of dynamic query filters with starfield displays. In: Proc. SIGCHI conference on Human factors in computing systems, pp. 313–317. ACM (1994)
5. Ahmed, Z., Yost, P., McGovern, A., Weaver, C.: Steerable clustering for visual analysis of ecosystems. In: EuroVA International Workshop on Visual Analytics (2011)
6. Andrienko, G., Andrienko, N.: Interactive cluster analysis of diverse types of spatio-temporal data. *SIGKDD Explor. Newsl.* **11**, 19–28 (2010)
7. Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., Giannotti, F.: Interactive visual clustering of large collections of trajectories. In: Proceedings of IEEE Symposium on Visual Analytics Science and Technology, pp. 3–10 (2009)
8. Andrienko, G., Andrienko, N., Wrobel, S.: Visual analytics tools for analysis of movement data. *SIGKDD Explorations* **9**(2), 38–46 (2007)
9. Andrienko, N., Andrienko, G.: Designing visual analytics methods for massive collections of movement data. *Cartographica* **42**(2), 117–138 (2007)
10. Bernard, J., Brase, J., Fellner, D., Koepler, O., Kohlhammer, J., Ruppert, T., Schreck, T., Sens, I.: A visual digital library approach for time-oriented scientific primary data. *Research and Advanced Technology for Digital Libraries* pp. 352–363 (2010)
11. Berndt, R., Blümel, I., Krottmaier, H., Wessel, R., Schreck, T.: Demonstration of user interfaces for querying in 3D architectural content in PROBADO3D. In: European Conference on Digital Libraries, *Lecture Notes in Computer Science*, vol. 5714, pp. 491–492. Springer (2009)
12. Bremm, S., von Landesberger, T., Andrienko, G., Andrienko, N.: Interactive analysis of object group changes over time. In: EuroVA International Workshop on Visual Analytics (2011)
13. Bremm, S., Landesberger, T.v., Bernard, J., Schreck, T.: Assisted descriptor selection based on visual comparative data analysis. *Computer Graphics Forum* **30**(3), 891–900 (2011)

14. Card, S.C., Mackinlay, J., Shneiderman, B.: *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers (1999)
15. Cui, W., Zhou, H., Qu, H., Wong, P.C., Li, X.: Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics* **14**(6), 1277–1284 (2008)
16. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)* **40**(2), 5 (2008)
17. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. *Information Retrieval* **11**(2), 77–107 (2008)
18. Frank, A., Asuncion, A.: UCI machine learning repository (2010). URL <http://archive.ics.uci.edu/ml>
19. Grinstein, G., Plaisant, C., Laskowski, S., O'connell, T., Scholtz, J., Whiting, M.: VAST 2008 Challenge: Introducing mini-challenges. In: *IEEE Symp. on Visual Analytics Science and Technology*, pp. 195–196 (2008)
20. Ivanov, Y., Wren, C., Sorokin, A., Kaur, I.: Visualizing the history of living spaces. *Transactions on Visualization and Computer Graphics* **13**(6), 1153–1160 (2007)
21. Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. (eds.): *Mastering the Information Age - Solving Problems with Visual Analytics*. Eurographics (2010)
22. Keim, D., Kohlhammer, J., May, T., Tomas, J.: Event summary of the workshop on visual analytics. *Computers & Graphics* **30**(2), 284–286 (2006)
23. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: *Visual Analytics: Scope and Challenges*. In: S. Simoff, M.H. Boehlen, A. Mazeika (eds.) *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Springer (2008). *Lecture Notes in Computer Science (LNCS)*
24. Kohonen, T.: *Self-Organizing Maps*, 3rd edn. Springer, Berlin (2001)
25. Kosara, R., Bendix, F., Hauser, H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics* pp. 558–568 (2006)
26. von Landesberger, T., Goerner, M., Schreck, T.: Visual analysis of graphs with multiple connected components. In: *Proceedings of IEEE Symposium on Visual Analytics Science and Technology* (2009)
27. Landesberger, T.v., Bremm, S., Bernard, J., Schreck, T.: Smart query definition for content-based search in large sets of graphs. In: *EuroVAST 2010*, pp. 7–12. European Association for Computer Graphics (Eurographics), Eurographics Association, Goslar (2010)
28. PANGAEA Publishing Network for Geoscientific & Environmental Data: <http://www.pangaea.de/>
29. Pelekis, N., Kopanakis, I., Marketos, G., Ntoutsi, I., Andrienko, G., Theodoridis, Y.: Similarity search in trajectory databases. In: *Proceedings of International Symposium on Temporal Representation and Reasoning*, pp. 129–140 (2007)
30. Rüger, S.: *Multimedia information retrieval. Synthesis Lectures on Information Concepts, Retrieval and Services*. Morgan & Claypool Publishers (2010)
31. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: *Shape Modeling Applications, 2004. Proceedings*, pp. 167–178. IEEE (2005)
32. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: *IEEE Visual Languages*, pp. 336–343 (1996)
33. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: *Proc. ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330. ACM Press (2006)
34. Tangelder, J.W.H., Velthkamp, R.C.: A survey of content based 3d shape retrieval methods. *Multimedia Tools Appl.* **39**(3), 441–471 (2008)
35. Thomas, J., Cook, K.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society (2005)
36. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents. In: *Proc. IEEE Symposium on Information Visualization*, pp. 51–58 (1995)
37. Yang, C., Chen, H., Hong, K.: Visualization of large category map for internet browsing. *Decision Support Systems* **35**(1), 89–102 (2003)