Visual Analytics Challenges

Daniel A. KEIM[1], Joern KOHLHAMMER[2] , Giuseppe SANTUCCI[3], Florian MANSMANN[1], Franz WANNER[1], Matthias SCHAEFER[1]

[1]*University of Konstanz, Universitaetsstr. 10, Konstanz, 78457, Germany*
Tel: +49 7531 88-3161, Fax: +49 7531 3062,
Email: {Daniel.Keim, Franz.Wanner, Florian.Mansmann}@uni-konstanz.de
[2]*Fraunhofer IGD, Fraunhoferstr. 5, Darmstadt, 64283, Germany*
Tel: +49 6151 155-646, Fax: +49 6151 155-139,
Email: Joern.Kohlhammer@igd.fraunhofer.de
[3]*Sapenzia Università di Roma, Via Ariosto 25, Roma, 00185, Italy*
Tel: +39 06 77274 006, Email: santucci@dis.uniroma1.it

Abstract: One of the most important challenges of the emerging Information Age is to effectively utilize the immense wealth of data and information acquired, computed and stored by modern information systems. On the one hand, the intelligent use of available data volumes and information extracted thereof offers large potential to realize technological progress and business success. On the other hand, there exists the severe danger that users and analysts get lost in irrelevant, or inappropriately processed or presented information, a problem, which is generally called the information overload problem. Visual Analytics is an emerging research discipline developing technology to make the best possible use of huge information loads in a wide variety of applications. The basic idea is to appropriately combine the strengths of intelligent automatic data analysis with the visual perception and analysis capabilities of the human user. In this paper, we will outline the most important challenges of the young research field.

1. Introduction

We live in a world, which faces a rapidly increasing amount of data to be dealt with on a daily basis. In the last decade, the steady improvement of data storage devices and means to create and collect data along the way influenced our way of dealing with information: Most of the time, data is stored without filtering and refinement for later use. Virtually every branch of industry or business, and any political or personal activity nowadays generate vast amounts of data. Making matters worse, the possibilities to collect and store data increase at a faster rate than our ability to use it for making decisions. However, raw data has no value in itself; instead we want to extract the information contained in it.

The **information overload problem** refers to the danger of getting lost in data, which may be a) irrelevant to the current task at hand, b) processed in an inappropriate way, or c) presented in an inappropriate way. Due to information overload, time and money are wasted, scientific and industrial opportunities are lost because we still lack the ability to deal with the enormous data volumes properly. People in both their business and private lives, decision-makers, analysts, engineers, emergency response teams alike, are often confronted with massive amounts of disparate, conflicting and dynamic information, which are available from multiple heterogeneous sources. There is a need for methods to effectively exploit and use the hidden information resting in unexplored data sources.

In many application areas success depends on the right information being available at the right time. Nowadays, the acquisition of raw data is no longer the driving problem: it is the ability to identify methods and models, which can turn the data into reliable and comprehensible knowledge. Any technology, that claims to overcome the information overload problem, has to provide answers for the following problems:

- Who or what defines the "relevance of information" for a given task?
- How can inappropriate procedures in a complex decision making process be identified?
- How can the resulting information be presented in a decision- or task-oriented way?

With every new "real-life" application, procedures are put to the test possibly under circumstances completely different from the ones under which they have been established. The awareness of the problem how to understand and analyse our data has been greatly increased in the last decade. Even as we implement more powerful tools for automated data analysis, we still face the problem of understanding and "analysing our analyses" in the future: Fully-automated search, filter and analysis only work reliably for well-defined and well-understood problems. The path from data to decision is typically quite complex. Even as fully-automated data processing methods represent the knowledge of their creators, they lack the ability to communicate their knowledge. This ability is crucial: if decisions that emerge from the results of these methods turn out to be wrong, it is especially important to examine the procedures.

The overarching driving vision of **Visual Analytics** [1][2][3] is to turn the information overload into an opportunity: just as *information visualisation* has changed our view on databases, the goal of Visual Analytics is to make *our way of processing* data and information transparent for an analytic discourse. The visualisation of these processes will provide the means of communicating about them, instead of being left with the results. Visual Analytics will foster the constructive evaluation, correction and rapid improvement of our processes and models and – ultimately – the improvement of our knowledge and our decisions.

On a grand scale, Visual Analytics provides technology that combines the strengths of human and electronic data processing. Visualisation becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective distinct capabilities for the most effective results. The user has to be the ultimate authority in giving the direction of the analysis along his or her specific task. At the same time, the system has to provide effective means of interaction to concentrate on this specific task since in many applications different people work along the path from data to decision. A visual representation will sketch this path and provide a reference for their collaboration across different tasks and abstraction levels.

Because Visual Analytics is an integrating discipline, application specific research areas contribute with existing procedures and models. Emerging from highly application-oriented research, dispersed research communities worked on specific solutions using the repertoire and standards of their specific fields. The requirements of Visual Analytics introduce new dependencies between these fields. This paper will map different perspectives onto a set of well established and agreed upon concepts and theories, allowing any scientific breakthrough in a single discipline to have a potential impact on Visual Analytics. In return, combining and lifting these multiple technologies onto a new general level will have a great impact on a large number of application domains.

2. Core Disciplines of Visual Analytics and their Challenges

Visualization

Visualization has emerged as a new research discipline during the last two decades. It can be broadly classified into *Scientific* and *Information* Visualization. In Scientific Visualization, the data entities to be visualized are typically 3D geometries or can be understood as scalar, vectorial, or tensorial fields with explicit references to time and space. A survey of current visualization techniques can be found in [4][5][6]. Often, 3D scalar fields are visualized using isosurfaces or semi-transparent point clouds (direct volume

rendering)[7]. To this end, methods based on optical emission- or absorption models are used which visualize the volume by ray-tracing or projection. Also, in the recent years significant work focused on the visualization of complex 3-dimensional flow data relevant e.g., in aerospace engineering[8]. While current research has focused mainly on efficiency of the visualization techniques to enable interactive exploration, more and more methods to automatically derive relevant visualization parameters come into focus of research. Also, interaction techniques such as focus & context[9] gain importance in scientific visualization.
In the last decade, Information Visualization has developed methods for the visualization of abstract data where no explicit spatial references are given[10][11][12]. In many application areas, the typically huge volumes of data require the appropriate usage of automatic data analysis techniques such as clustering or classification as preprocessing prior to visualization. Research in this direction is currently emerging.

Data Management

An efficient management of data of various types and qualities is a key component of Visual Analytics as this technology typically provides the input of the data, which are to be analyzed. Generally, a necessary precondition to perform any kind of data analysis is an integrated and consistent data basis[13][14]. Database research has until the last decade focused mainly on aspects of efficiency and scalability of exact queries on homogeneous, structured data. With the advent of the Internet and the easy access it provides to all kinds of heterogeneous data sources, the database research focus has shifted towards integration of heterogeneous data. Finding integrated representations for different data types such as numeric data, graphs, text, audio and video signals, semi-structured data, semantic representations and so on is a key problem of modern database technology. But the availability of heterogeneous data not only requires the mapping of database schemata but includes also the cleaning and harmonization of uncertainty and missing data in the volumes of heterogeneous data. Modern applications require such intelligent data fusion to be feasible in near real-time and as automatically as possible[15]. New forms of information sources such as data streams[16], sensor networks[17] or automatic extraction of information from large document collections (e.g., text, HTML) result in a difficult data analysis problem which to support is currently in the focus of database research[18].

Data Mining and Analysis

Data Mining and Analysis researches methods to automatically extract valuable information from raw data by means of automatic analysis algorithms[19][20][21] and to approve existing models about the data. Approaches developed in this area can be best described by the addressed analysis tasks. A prominent such task refers to learning from supervised learning from examples: Based on a set of training samples, deterministic or probabilistic algorithms are used to learn models for the classification (or prediction) of previously unseen data samples[22]. A huge number of algorithms have been developed to this end such as Decision Trees, Support Vector Machines, Neuronal Networks, and so on. A second prominent analysis task is that of cluster analysis[23][24], which aims to extract structure from data without prior knowledge being available. Solutions in this class are employed to automatically group data instances into classes based on mutual similarity, and to identify outliers in noisy data during data preprocessing for subsequent analysis steps. Further data analysis tasks include tasks such as association rule mining (analysis of co-occurrence of data items) and dimensionality reduction. While data analysis initially was developed for structured data, recent research aims at analyzing also semi-structured and complex data types such as web documents or multimedia data[25].

It has recently been recognized that Visualization and Interaction are highly beneficial in arriving at optimal analysis results. In almost all data analysis algorithms a variety of parameters needs to be specified, a problem which is usually not trivial and often needs supervision by a human expert. Visualization is also a suitable means for appropriately communicating the results of the automatic analysis, which often is given in abstract representation, e.g., a decision tree. Visual Data Mining methods[26] try to achieve this.

Spatio-Temporal Data Analysis

While many different data types exist, one of the most prominent and ubiquitous data types is data with references time and space. The importance of this data type has been recognized by a research community, which formed around spatio-temporal data management and analysis[27]. In *geospatial* data research, data with references in the real world coming from e.g., geographic measurements, GPS position data, remote sensing applications, and so on is considered. Finding spatial relationships and patterns among this data is of special interest, requiring the development of appropriate management, representation and analysis functions and visualization often plays a key role in the successful analysis of geospatial data[28][29].

The analysis of data with references both in space and in time is a challenging research topic. Major research challenges include scale and uncertainty of the data. Regarding scale, clusters and other phenomena may only occur at particular scales, which may not be the scale at which data is recorded. So in addition to needing scalable techniques, research here needs to deal with the effects of considering spatio-temporal data at different scales. Regarding data uncertainty, spatio-temporal data are often incomplete, interpolated, collected at different times, based upon different assumptions etc.

Perception and Cognition

Effective utilization of the powerful human perception system for visual analysis tasks requires the careful design of appropriate human-computer interfaces. Psychology, Sociology, Neurosciences and Design each contribute valuable results to the implementation of effective visual information systems. Research in this area focuses on user-centred analysis and modelling (Requirement Engineering), the development of principles, methods and tools for design of perception-driven, multimodal interaction techniques for visualization and exploration of large information spaces, as well as usability evaluation of such systems[30][31][32]. On the technical side, research in this area is influenced by two main factors: (1.) The availability of improved display resources (hardware), and (2.) Development of novel interaction algorithms incorporating machine recognition of the actual user intent and appropriate adaptation of main display parameters such as the level of detail, data selection, etc. by which the data is presented. Important problems addressed in this area include the research of perceptual, cognitive and graphical principles which in combination lead to improved visual communication of data and analysis results.

Evaluation

The above described research disciplines require cross-discipline support regarding the evaluation of the found solutions, and need certain infrastructure and standardization grounding to build on effectively. In the field of information visualization, standardization and evaluation came into the focus of research only recently. It has been realized that a general understanding of the taxonomies regarding the main data types and user tasks[33] to be supported are highly desirable for shaping Visual Analytics research.

How to assess (evaluate) the value of visualization is a topic of lively debate[34][35]. A common ground that can be used to position and compare future developments in the field of data analysis is needed since a more rigorous and overall scientific perspective will lead to a more effective and efficient development of innovative methods and techniques.

3. Application Areas of Visual Analytics

Visual Analytics is a highly application oriented discipline driven by practical requirements. It is essential in a wide range of application areas where large information spaces have to be processed and analyzed.

Public Safety & Security is one important application area where Visual Analytics may contribute with advanced solutions. Analysts need to constantly monitor huge amounts of heterogeneous information streams, correlating information of varying degrees of abstraction and reliability, assessing the current level of public safety, triggering alert in case of alarming situations being detected. Data integration and correlation combined with appropriate analysis and interactive visualization is promising to develop more efficient tools for the analysis in this area.

The study of Environment and Climate change often requires the examination of long-term weather records and logs of various sensors, in a search for patterns that can be related to observations such as changes in animal populations, or in meteorological and climatic processes for instance. These requirements call for the development of systems allowing visual and graphical access to historical monitoring data and predictions from various models in search for or in order to validate patterns building over time.

A major field in the area of Visual Analytics covers physics and astronomy, including applications like flow visualization, fluid dynamics, molecular dynamics, nuclear science and astrophysics, to name just a few of them. By common data analysis techniques like knowledge discovery, astronomers can find new phenomena, relationships and useful knowledge about the universe, but although a lot of the data only consists of noise, a Visual Analytics approach can help separating relevant data from noise and help identifying unexpected phenomena inside the massive and dynamic data streams. One example for a Visual Analytics application is the simulation of a Supernova. The SciDAC program has brought together tremendous scientific expertise and computing resources within the Terascale Supernova Initiative (TSI) project to realize the promise of terascale computing for attempting to answer some of the involved questions[36].

Another major field in the area of Visual Analytics covers business applications. The financial market with its thousands of different stocks, bonds, futures, commodities, market indices and currencies generates a lot of data every second, which accumulates to high data volumes throughout the years. The main challenge in this area lies in analyzing the data under multiple perspectives and assumptions to understand historical and current situations, and then monitoring the market to forecast trends and to identify recurring situations. Visual Analytics applications can help analysts obtaining insights and understanding into previous stock market development, as well as supporting the decision making progress by monitoring the stock market in real-time in order to take necessary actions for a competitive advantage, with powerful means that reach far beyond the numeric technical chart analysis indicators or traditional line charts. One popular application in this field is the well-known Smartmoney[37], which gives an instant visual overview of the development of the stock market in particular sectors for a user-definable time frame.

"Financial matrix"[38] visualisation is a relevance driven visualization technique of financial performance measures. Standard statistical measures for technical financial data analysis often produce insufficient and misleading results that do not reflect the real performance of an asset. The technique for visualizing financial time series data eliminates these inadequacies, offering a complete view on the real performance of an asset. The technique is enhanced by relevance and weighting functions according to the users' preferences in order to emphasize specific regions of interest. Figure 1 shows an example of the visualization technique; the performance of funds is highlighted green or red over various possible buying and selling points over the time.

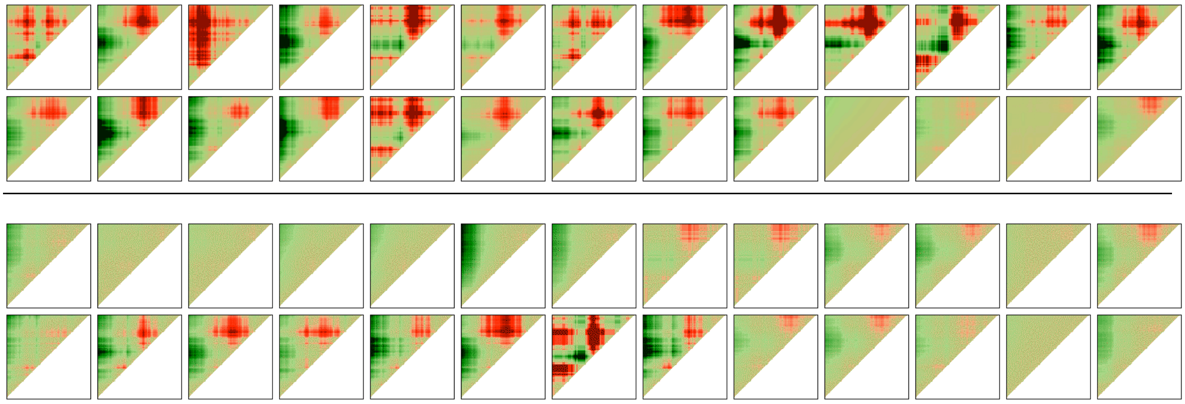


Fig. 1: Funds performance analysis - Top: 26 funds of the same financial institution and their performance over time, green means plus, red minus; good and bad performing funds are easily distinguishable. Bottom: The 26 funds of another financial institution and their performance over time. Funds of the financial institution shown below are more continuous and have less volatility.

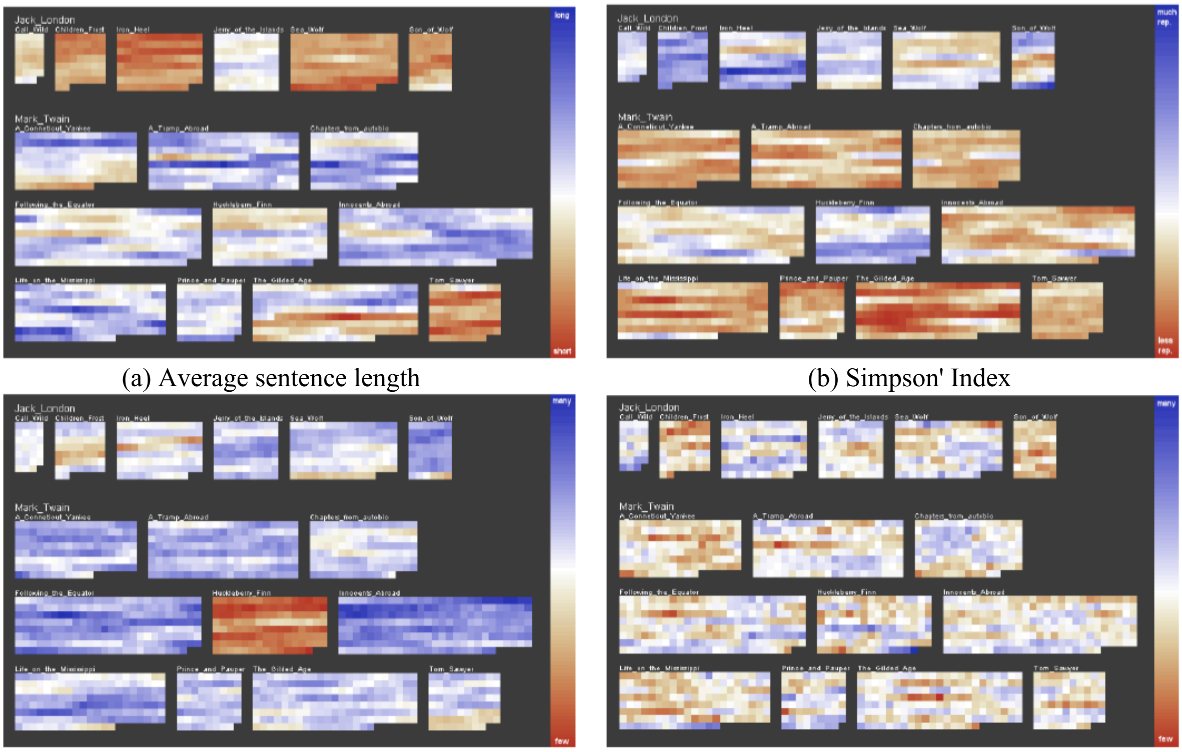
In computer-based literary analysis different types of features are used to characterize a text. Usually, only a single feature value or vector is calculated for the whole text. "Literature fingerprinting"[39] combines automatic literature analysis methods with an effective visualization technique to analyze the behaviour of the feature values across the text. For an interactive visual analysis, a sequence of feature values per text is calculated and presented to the user as a characteristic fingerprint. The feature values may be calculated on different hierarchy levels, allowing the analysis to be done on different resolution levels. Figure 2 gives an impression of the technique.

4. Conclusions

Visual Analytics combines strengths from information analytics, geospatial analytics, scientific analytics, statistical analytics, knowledge discovery, data management & knowledge representation, presentation, production & dissemination, cognition, perception, and interaction. It is a goal-oriented process to gain insight into heterogeneous, contradictory and incomplete data through the combination of automatic analysis methods with human background knowledge and intuition.

Unlike described in the information seeking mantra ("overview first, zoom/filter, details on demand"[40]) the Visual Analytics process comprises the application of automatic analysis methods before and after the interactive visual representation is used since current and especially future data sets are complex and too large to be visualized in a straightforward manner. Therefore, we present the Visual Analytics mantra:

"Analyse First -
Show the Important -
Zoom, Filter and Analyse Further -
Details on Demand" [3]



(a) Average sentence length

(b) Simpson' Index

(c) Hapax Legomena

(d) Hapax Dislegomena

Fig. 2: Analysis of the discrimination power of several text measures for authorship attribution [39]. Each pixel represents a text block, which are grouped into books. Color is mapped to the feature value. If a measure is able to discriminate between the two authors, the books in the first line (written by J. London) are visually set apart from the remaining books (written by M. Twain). While measures a) and b) are discriminative, d) is not.

The Visual Analytics mantra could be exemplarily applied in the context of data analysis for network security. Visualizing the raw data is unfeasible and rarely reveals any insight. Therefore, the data is first analysed (i.e., compute changes, intrusion detection analysis, etc.) and then displayed. The analyst proceeds by choosing a small suspicious subset of the recorded intrusion incidents by applying filters and zoom operations. Finally, this subset is used for a more careful analysis. Insight is gained in the course of the whole Visual Analytics process.

Acknowledgement

This work was funded by the VisMaster Coordination Action in the Future and Emerging Technologies (FET) programme under FET-Open grant number 225429.

References

[1] J.J. Thomas and K.A. Cook, Illuminating the Path: The Research and Development Agenda for Visual Analytics, IEEE CS Press, 2005.

[2] D. A. Keim, G. Andrienko, J.-d. Fekete, Carsten Gorg, Jorn Kohlhammer, G. Melancon: Visual Analytics: Definition, Process, and Challenges, 2007, Dagstuhl Group Report.

[3] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Hartmut Ziegler: Challenges in Visual Data Analysis, Information Visualization (IV 2006), Invited Paper, July 5-7, London, United Kingdom, IEEE Press, 2006.

[4] C. Johnson, C. Hanson (Eds.): Visualization Handbook, Kolam Publishing, 2004.

[5] H. Schumann., W. Müller: Visualisierung - Grundlagen und allgemeine Methoden; Springer 2000

[6] D. Keim, T. Ertl (Hrsg.): Scientific Visualization (in German), Information Technology, Oldenbourg. 46(3), 2004

[7] K. Engel, M. Hadwiger, J. M. Kniss, C. Rezk-Salama, D. Weiskopf: Real-Time Volume Graphics, AK Peters, 2006.

[8] X. Tricoche, G. Scheuermann, H. Hagen: Tensor Topology Tracking: A Visualization Method For Time-Dependent 2D Symmetric Tensor Fields, Computer Graphics Forum, Proceedings Eurographics, 20(3), 2001

[9] J. Krüger, J. Schneider, R. Westermann: ClearView: An Interactive Context Preserving Hotspot Visualization Technique, IEEE TVCG, 12(5):941-948, 2006.

[10] R. Spence: Information Visualization, ACM-Press, 2000

[11] S. K. Card, J. D. Mackinlay, and B. Shneiderman: Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann Publishers, 1999

[12] D. Keim, M. Ward: Visual Data Mining Techniques, Book Chapter in: Intelligent Data Analysis, an Introduction, 2002

[13] J. Han, M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.

[14] D. Hand, H. Mannila, and P. Smyth: Principles of Data Mining, MIT Press, August 2001.

[15] F. Naumann, A. Bilke, J. Bleiholder, M. Weis: Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies. IEEE Data Eng. Bull. 29(2): 21-31 (2006).

[16] A. Das, J. Gehrke, M. Riedewald: Semantic Approximation of Data Stream Joins. IEEE Trans. Knowl. Data Eng. 17(1): 44-59 (2005).

[17] A. Meliou, D. Chu, J. M. Hellerstein, C. Guestrin, W. Hong: Data gathering tours in sensor networks. IPSN 2006.

[18] J. Widom: Trio: A System for Integrated Management of Data, Accuracy, and Lineage. CIDR 2005: 262-276.

[19] Oded Maimon, Lior Rokach (Eds.): The Data Mining and Knowledge Discovery Handbook. Springer 2005.

[20] M. Ester, J. Sander: Knowledge Discovery in Databases: Techniken und Anwendungen, Springer, 2000.

[21] T. M. Mitchell: Machine Learning, Berkeley, Calif. McGraw-Hill, 1997.

[22] Duda, Hart, Stork: Pattern Classification. Wiley & Sons, 2000.

[23] J. Han, M. Kamber: Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000

[24] D. Hand, H. Mannila, and P. Smyth: Principles of Data Mining, MIT Press, August 2001.

[25] Perner: Data Mining on Multimedia Data, Springer 2002.

[26] D. Keim, M. Ward: Visual Data Mining Techniques, Book Chapter in: Intelligent Data Analysis, an Introduction by D. Hand and M. Berthold, 2nd Edition, 2002, Springer.

[27] Andrienko, Andrienko: Exploratory Analysis of Spatial and Temporal Data. A Systematic Approach. Springer 2005

[28] J. Dykes, A.M. MacEachren, M.-J. Kraak, Exploring geovisualization, 2005, Elsevier

[29] D. A. Keim, S. C. North, C. Panse, M. Sips: Pixel based Visual Mining of Geo-Spatial Data, Computers & Graphics (CG&A), Vol. 28, No. 3, pp. 327-344, Elsevier Science, June, 2004

[30] J. A. Jacko and A. Sears (Eds.). The Handbook for Human Computer Interaction. Mahwah: Lawrence Erlbaum & Associates, 2002.

[31] A. Dix, J. Finlay, G. D. Abowd, and R. Beale: Human-Computer Interaction. Prentice Hall, 2003.

[32] B. Shneiderman and C. Plaisant: Designing the User Interface: Strategies for Effective Human-Computer Interaction. 4th ed. Addison Wesley, 2004.

[33] Robert Amar, James Eagan, John Stasko: Low-Level Components of Analytic Activity in Information Visualization, Proc IEEE Symposium on Information Visualization, 2005.

[34] J.J. van Wijk. The Value of Visualization. In Proceedings IEEE Visualization 2005, pp 79-86, 2005.

[35] C. North. Toward Measuring Visualization Insight. IEEE Computer Graphics and Applications 26(3), pages 6-9, 2006.

[36] K.-L. Ma, E. Lum, H. Yu, H. Akiba, M.-Y. Huang, Y. Wang, and G. Schussman, \Scientific discovery through advanced visualization," in Proceedings of DOE Sci-DAC 2005 Conference, San Francisco, June 2005.

[37] M. Wattenberg, "Visualizing the stock market," in CHI '99: CHI '99 extended abstracts on Human factors in computing systems. New York, NY, USA: ACM Press, 1999, pp. 188-189.

[38] H. Ziegler, T. Nietzschmann, D.A. Keim: Relevance Driven Visualization of Financial Performance Measures, EuroVis 2007: Eurographics/IEEE-VGTC Symposium on Visualization, Stockholm, Sweden, 23 - 25 May, 2007.

[39] D.A. Keim, D. Oelke: Literature Fingerprinting: A New Method for Visual Literary Analysis, IEEE Symposium on Visual Analytics and Technology (VAST 2007), 2007.

[40] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In IEEE Symposium on Visual Languages, pages 336–343, 1996.