Informationsvisualisierung

Methoden und Perspektiven

Heidrun Schumann
Universität Rostock, Institut für Informatik
schumann@informatik.uni-rostock.de

Abstract: Ziel der Informationsvisualisierung ist es, abstrakte Daten graphisch so zu repräsentieren, dass strukturelle Zusammenhänge und relevante Eigenschaften intuitiv erfasst werden können. Damit soll die interaktive Exploration komplexer Datenmengen unterstützt werden. Aktuelle Themen adressieren vor allem Komplexität und Umfang heutiger Datensätze, eine stärkere Berücksichtigung des Anwenders sowie die Verknüpfung von visuellen und automatischen Methoden. Im Paper werden sowohl grundlegende Herangehensweisen umrissen als auch konkrete Methoden zur Visualisierung von Datenwerten und hierarchischen Strukturen vorgestellt. Abschließend wird ein kurzer Ausblick auf künftige Entwicklungen gegeben.

1. Grundlegende Konzepte der Informationsvisualisierung

Die Informationsvisualisierung ist seit den 90er Jahren als eigenständiges Wissensgebiet etabliert. Sie untersucht die graphische Repräsentation abstrakter Daten, die nicht notwendigerweise einen physikalischen Bezug aufweisen (vgl. [1]). Ziel ist es, die große Leistungsfähigkeit des menschlichen visuellen Systems auszunutzen, um die charakteristischen Eigenschaften einer Datenmenge zu erfassen. Um dieses Ziel zu erreichen, hat Shneiderman das *Information Seeking Mantra* aufgestellt: „Overview first, Zoom and Filter, then Detail on Demand“. Hiermit wird ein prinzipielles Vorgehen der visuellen Analyse beschrieben. Ausgangspunkt ist demnach ein Überblicksbild (*Overview*), das allgemeine Eigenschaften der Datenmenge geeignet präsentiert und als Ausgangspunkt für weitere Untersuchungen dient. Mittels *Zooming* werden anschließend bestimmte Bildbereiche vergrößert und so der Fokus auf interessierende Teilbereiche gelenkt. Das *Filtering* erlaubt es zudem, Informationen auszublenden, die in einem gegebenen Kontext nicht zur Problemlösung beitragen. Für eine tiefer gehende Analyse lassen sich dann verschiedene Details anzeigen (*Details on Demand*). Dieses prinzipielle Vorgehen schließt 3 Aspekte ein, welche die Informationsvisualisierung charakterisieren:

- Es werden unterschiedliche Bilder erzeugt, um unterschiedliche Sichten auf die Daten zu ermöglichen.
- Zum Explorieren und Wechseln der Sichten stehen vielfältige Interaktionstechniken bereit.
- Für das Berechnen allgemeiner Eigenschaften, das Ausblenden nicht-relevanter Informationen sowie die Ergänzung von Details werden automatische Analyseverfahren (z.B. aus der Statistik) eingesetzt.

Die Informationsvisualisierung ist also ein in hohem Maße interaktiver Prozess, der visuelle und automatische Methoden verknüpft. Dieser Zusammenhang wird durch das *Data State Reference Model* von Chi [2] beschrieben (vgl. Abbildung 1).

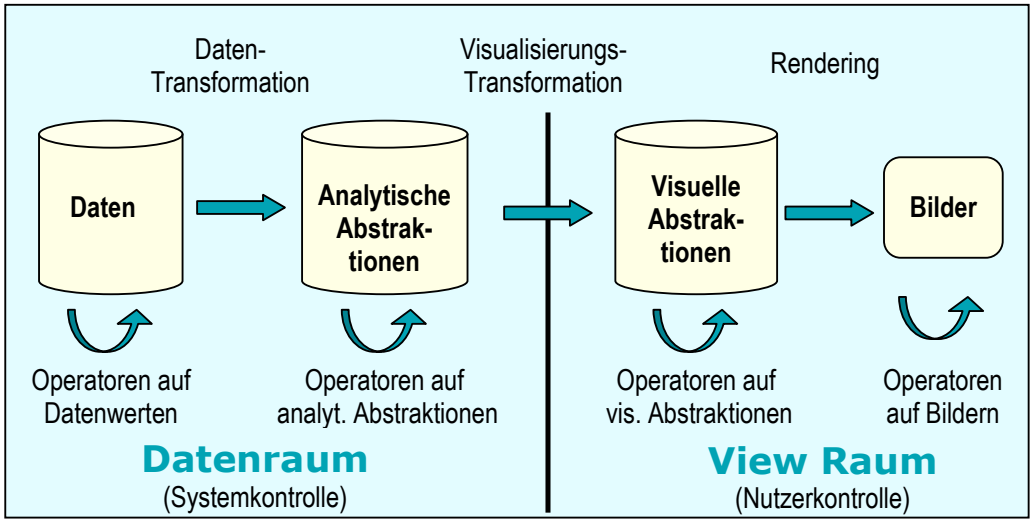

Abbildung 1: Data State Reference Model [2]

Dieses Modell systematisiert zwei Aspekte:

- Die stufenweise Abbildung der Daten über analytische und geometrische Abstraktionen auf Bilddaten,
- die Anwendung vielfältiger Operatoren auf den unterschiedlichen Stufen des Modells.

Die Verknüpfung der Operatoren führt zu einer datenfluß-orientierten Beschreibung des Visualisierungsprozesses, die den meisten Visualisierungssystemen zugrunde liegt. Dabei sei angemerkt, dass heutige Systeme nicht dieselbe Unterstützung auf allen Stufen des Data State Reference Modells anbieten. Die einzelnen Systeme unterscheiden sich vor allem dadurch, welche konkreten Operatoren sie bereitstellen.

Eine wichtige Anforderung an den Visualisierungsprozess und damit an heutige Visualisierungssysteme ist die Skalierbarkeit. In [3] wird zwischen *Information Scalability*, *Visual Scalability*, *Display Scalability* und *User Scalability* unterschieden Hiermit wird sowohl die Bereitstellung als auch die Präsentation von Informationen auf verschiedenen Genauigkeitsstufen eingefordert, aber auch die Vielfalt der Nutzeranforderungen und Geräteeigenschaften adressiert. Im Folgenden sollen hierzu konkrete Beispiele vorgestellt werden.

2. Visualisierung von Datenwerten

Immer größere und komplexere Datenmengen stellen auch immer höhere Anforderungen an den Visualisierungsprozess, so dass man ohne Skalierung nicht mehr auskommt. Dies soll anhand von 3 Beispielen demonstriert werden:

Skalierung der Darstellung - Die Data Table View: Die *Table Lens* ist eine bekannte und weit verbreitete Methode zur Visualisierung von Tabellendaten [4]. Das Grundprinzip besteht darin, die Tabellenansicht mit einer Fokus & Kontext-Technik zu verknüpfen und damit eine Skalierung der Darstellung zu erreichen. Datenwerte im Kontext werden graphisch kodiert, so dass sich die entsprechenden Tabellenzeilen stark verkleinert anzeigen lassen. Dagegen werden die Datenwerte im Fokus numerisch ausgegeben. In [5] wird mit der Data Table View diese Vorgehensweise durch die Integration eines allgemeinen Sortierungsmechanismus auf der Basis *Selbstorganisierender Maps* (SOM) [6] erweitert. Dadurch werden Tabellenzeilen mit ähnlichen Informationen benachbart angeordnet, so dass sich Korrelationen gut erkennen

lassen. Abbildung 2 zeigt einen multi-variaten Auto-Datensatz mit 392 Tabellenzeilen. Die linke Abbildung zeigt keine Sortierung mit dem Fokus auf einer bestimmten Tabellenzeile. Die rechte Abbildung enthält eine SOM- Sortierung. Es ist gut zu erkennen, dass zwischen den ersten 4 Spalten Korrelationen bestehen.

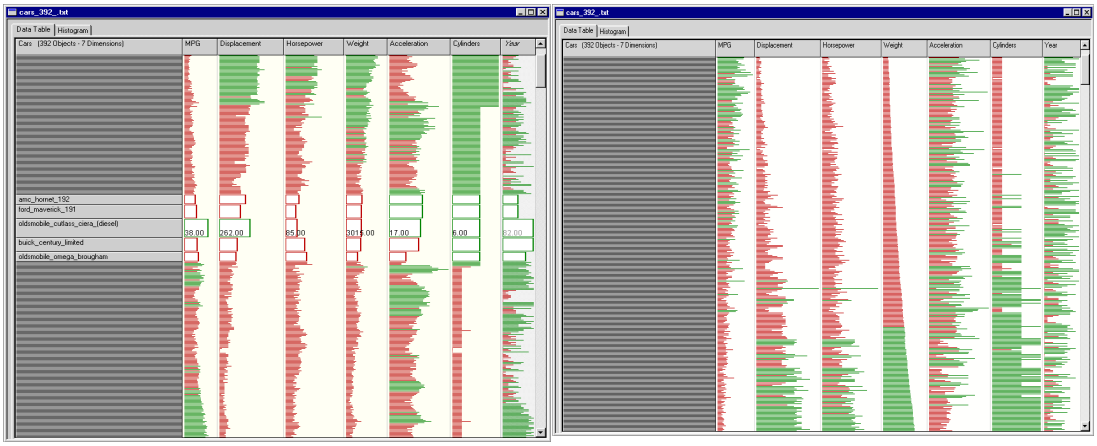

Abbildung 2: Darstellung eines Datensatzes mit der Data Table View [5], links ohne und rechts mit SOM-Sortierung, Werte über einem bestimmten Schwellwert sind grün, anderenfalls braun dargestellt

In [7] wird die graphische Kodierung durch Anwendung der *Two-Tone Pseudo-Coloring* Technik [8] weiter verbessert und außerdem ein hybrider Clusteralgorithmus integriert, um einerseits starke Schwankungen auszugleichen, andererseits aber auch strukturelle Eigenschaften der Daten zu kommunizieren. Abbildung 3 zeigt diese Vorgehensweise. In Abbildung 3, oben links, wird der multivariate Autodatensatz aus Abbildung 2 mit der Two-Tone-Pseudo-Farbkodierung unsortiert dargestellt. Abbildung 3, oben rechts, zeigt denselben Datensatz nach Anwendung eines hierarchischen Clusterings (die Cluster-Struktur ist als *Icicle Plot* in der letzten Tabellenspalte veranschaulicht). Die Daten eines ausgewählten Clusters sind blau markiert. Abbildung 3 unten veranschaulicht die interaktive Exploration der Clusterhierarchie. Nach und nach können bestimmte Zweige des *Icicle-Plots* ausgeklappt und so die zugehörigen Cluster analysiert werden.

Abbildung 3: Darstellung eines Datensatzes mit der Two Tone Table Lens [7], links oben: ohne Sortierung; rechts: Anwendung eines hierarchischen Clusteralsorithmus; unten: schrittweises Ausklappen von Clustern

Skalierung der Daten

Das *Data State Reference Model* aus Abbildung 1 weist die Bildung von analytischen Abstraktionen als eigenständigen Schritt aus und unterstützt damit die Skalierung im Datenraum. Üblicherweise werden hierfür sowohl Clusteralgorithmen als auch multi-dimensionale Skalierungsmethoden eingesetzt [9]. Insbesondere durch ein *Hierarchisches Clustern* lassen sich unterschiedliche Abstraktionsstufen der Daten erzeugen, die, je nach den gegebenen Anforderungen, für einen ersten Analyseschritt auf einem entsprechenden Datenlevel angezeigt werden können. Ein dynamischer Wechsel von einer Granularitätsstufe zu einer anderen ist dabei problemlos möglich. Das Beispiel der *Two Tone Table Lens* zeigt bereits, wie auf diese Weise die Skalierung im Darstellungsraum mit einer Skalierung im Datenraum verbunden werden kann. Die folgenden 2 Beispiele demonstrieren noch einmal, wie sich verschiedene Abstraktionsstufen der Daten in einer visuellen Repräsentation gemeinsam darstellen lassen.

- **Clusteranalyse und Visualisierung**:

  Die Calendar View aus [10] ist eine effektive Visualisierungstechnik, die gleichzeitig die Ergebnisse eines Clusteralgorithmus und die originalen Datenwerte kommuniziert. In Abbildung 4 ist dieses Vorgehen am Beispiel der Visualisierung von Klimadaten aus dem Potsdamer Institut für Klimafolgenforschung demonstriert [11]. Gezeigt werden die täglich und stundenweise gemessenen Temperaturwerte in Potsdam für das Jahr 2000. Die Kalenderdarstellung auf der rechten Seite zeigt typische Tagesmuster. Jeder Tag ist entsprechend der Zugehörigkeit zu einem Cluster eingefärbt. Die Liniengraphik auf der linken Seite zeigt dagegen die Temperaturfunktion für ausgewählte Cluster, Tage oder Monate.

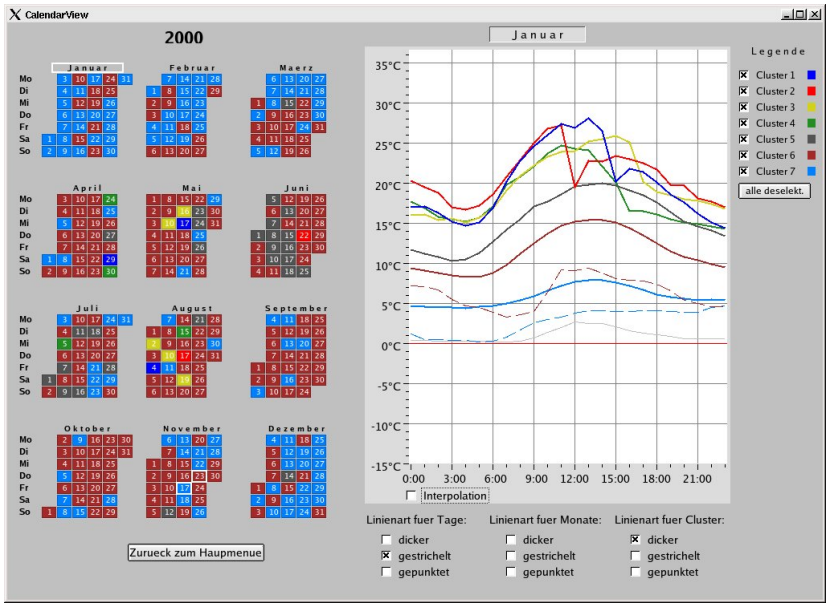

Abbildung 4: Visualisierung der geclusterten Temperaturdaten des PIK im Jahr 2000 mit der Calendar View aus [10]

- **PCA – Analyse und Visualisierung**:

  In der Visualisierung wird die Hauptkomponentenanalyse hauptsächlich als Vorverarbeitungsschritt zur Dimensionsreduktion eingesetzt (vgl. auch [9]). In [12] wird gezeigt, wie sich die Ergebnisse der PCA auf allen Stufen der Visualisierungspipeline einsetzen lassen, um expressive Darstellungen zu erzeugen. Insbesondere die gemeinsame Darstellung von PCA- und originalen Datenwerten eröffnet

neue Perspektiven für die visuelle Analyse. In Abbildung 5 sind die Werte eines demographischen Datensatzes gemeinsam mit den Ergebnissen der PCA- Analyse dargestellt.

Das linke obere Bild aus Abbildung 5 zeigt die Loadings in Kombination mit den Variablen des Datensatzes. Positive Werte sind blau, negative Werte gelb kodiert. Die Werte in einer Zeile zeigen den Einfluss einer Variablen auf die durch die PCA- Achsen repräsentierten Trends. Interessant ist hier der gegensätzliche Einfluss der Lebenserwartung der männlichen bzw. weiblichen Bevölkerung auf die PC9. In der rechten oberen Darstellung sind die Loadings bezogen auf die Signifikanzwerte skaliert. Hohe Werte in einer Spalte der Tabelle dienen als Indikator für eine hohe Relevanz der entsprechenden PCA- Achse. Das untere Bild zeigt die Scores in Verbindung mit den Variablen des Datensatzes. Die Darstellung wurde so skaliert, dass der Fokus auf Ausreißern liegt, das heißt auf Datensätzen, die hohe Werte für weniger relevante PCA- Achsen aufweisen.

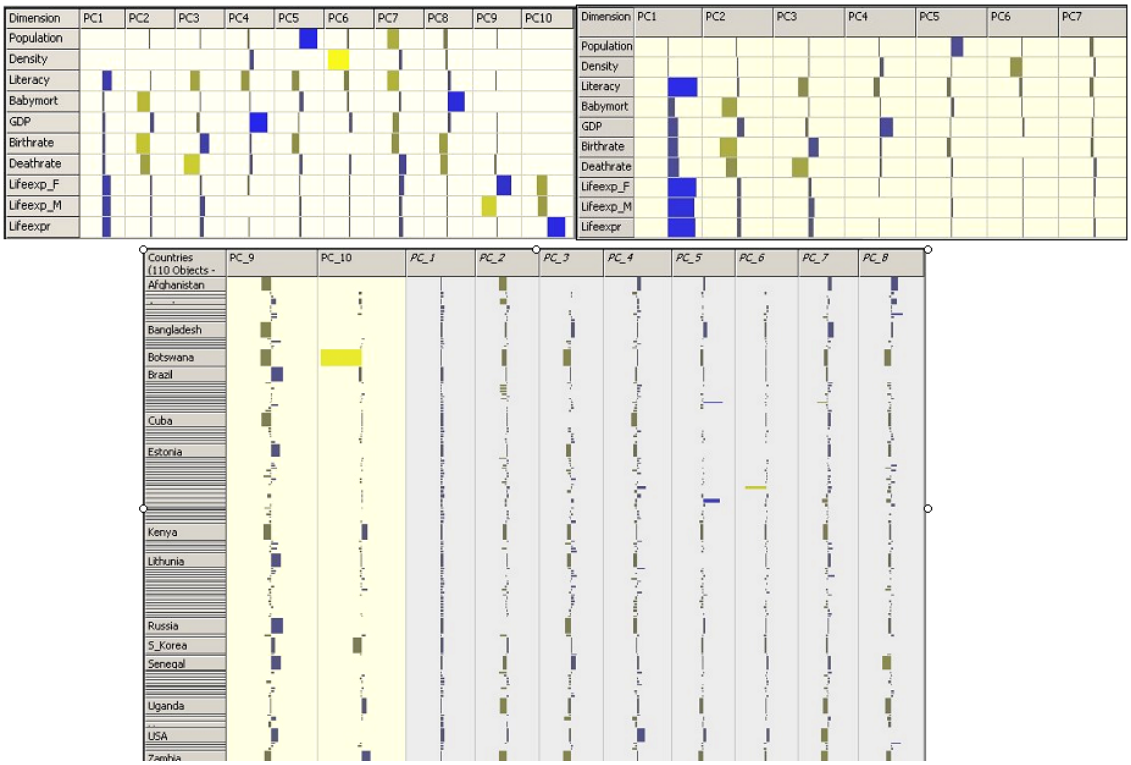

Abbildung 5: Data Table View zur kombinierten Visualisierung vonPCA-Daten und Originaldaten [12]; oben: Loadings und Variable (rechts normiert bezogen auf die Signifikanz); unten: Scores und Datensätze mit Fokus auf Ausreißer

3. Visualisierung von Strukturen

Neben der Exploration der Datenwerte ist es auch wichtig, Beziehungen zwischen den Variablen und Datensätzen zu analysieren. Hierzu werden die Datenobjekte als Knoten eines Graphen und ihre Beziehungen untereinander als Kanten aufgefasst. In der Visualisierung beschäftigt man sich vor allem mit der Darstellung hierarchischer Strukturen, weil hiermit die Forderung nach Skalierbarkeit sehr gut umgesetzt werden kann, z.B. durch das interaktive Ein- und Ausblenden von Teilbäumen oder Ebenen der Hierarchie. Zur Visualisierung von Hierarchien werden verschieden Techniken eingesetzt. Dabei lassen sich folgende prinzipielle Vorgehensweisen unterscheiden:

- 2D vs. 3D
  Das Layout der Knoten erfolgt in der Ebene oder im 3D-Raum

- Achsenparallel vs. Radial
  Das Layout der Knoten wird entweder entsprechend der Achsen des Präsentationsraumes ausgerichtet oder radial um den Wurzelknoten angeordnet.

- Explizit vs. Implizit
  Die Kanten werden explizit gezeichnet oder implizit durch entsprechende Anordnungen der Knoten veranschaulicht.

Üblicherweise werden heute, insbesondere für kleinere Knoten- und Kantenmengen, explizite Techniken, die so bezeichneten Node-Link-Diagramme eingesetzt. Allerdings werden mit zunehmender Datengröße auch implizite Techniken immer populärer. Eine bekannte und oft verwendete implizite Technik ist die *Treemap-Darstellung* [13]. Hierbei wird der Wurzelknoten durch ein umschreibendes Rechteck präsentiert. Das Prinzip besteht nun darin, dieses Rechteck rekursiv entsprechend der Anzahl der jeweiligen Kindknoten weiter zu unterteilen. Das letzte Unterteilungslevel zeigt die Blätter der Hierarchie. Es gibt viele verschiedene Spielarten der Treemap-Darstellung, die sich vor allem dadurch unterscheiden, welche Elemente für die fortlaufende Unterteilung genutzt werden und wie die konkrete Anordnung der Elemente erfolgt. Heutige kommerziellen Visualisierungs-Tools wie Spotfire oder IBM ILOG Elixir bieten Treemap- Visualisierungen an.

Abbildung 6 zeigt verschiedene Beispiele zur Visualisierung von Hierarchien; eine einfache Treemap-Darstellung ist unten links zu sehen.

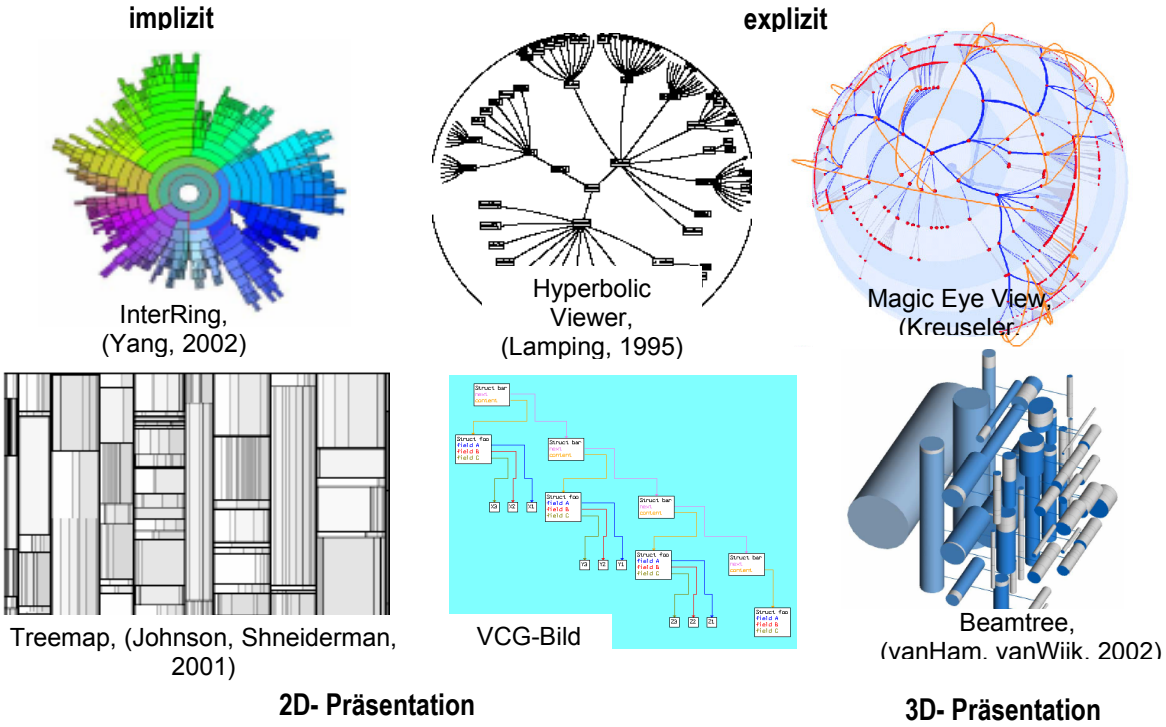

Abbildung 6: Beispiele zur Hierarchievisualisierung; oben : radiale Anordnungen, unten achsenparallele Anordnungen

4. Ausblick

Die interaktive Informationsvisualisierung stellt in unserer heutigen Informationsgesellschaft ein unverzichtbares Hilfsmittel zur Exploration großer Datenmengen dar. Nur durch Verknüpfung von visuellen und interaktiven Methoden kann Skalierbarkeit gewährleistet, und der Nutzer direkt in den Analyseprozess mit einbezogen werden. Aktuelle Entwicklungen gehen noch einen Schritt weiter. Das neue Forschungsgebiet des *Visual Analytics* [3] stellt eine weit reichende Verknüpfung von visuellen und automatischen Methoden aus verschiedenen Gebieten in den Fokus. Durch die Kombination unterschiedlicher Ansätze soll die Analyse extrem großer, komplexer und heterogener Datensätze umfassend unterstützt werden. Das EU-Projekt „VisMaster – Mastering the Information Age“ [14] greift diese Problematik auf. Ziel diese Netzwerkprojektes ist es, in Europa eine entsprechende Community zu bilden und mit einer Roadmap Herausforderungen, Möglichkeiten und offene Probleme für zukünftige Forschungsschwerpunkte zu formulieren.

Literatur

[1] Shneiderman, B.: The Eyes have it: A task by data type taxonomy of information visualizations. Proc. IEEE Symposium on Visual Languages'96, IEEE Computer Society Press Los Alamos, CA, (1996), pp. 336–343.

[2] Chi, E. H.: A Taxonomy of Visualization Techniques using the Data State Reference Model. Proc. IEEE Symposium on Information Visualization (InfoVis 2000), IEEE Press (2000), S. 69–75.

[3] Thomas, J.J.; Cook, K.A. (eds): Illuminating the Path: The Research and Development Agenda for Visual Analytics, NVAC, IEEE Computer Society Press, 2005

[4] Rao, R.; Card, S.: The Table Lens: Merging Graphical and Symbolical Representations in an Interactive Focus+Context Visualization for Tabular Information. Proceedings “Human Factors in Computing Systems, Apr. 1994, S. 318-322

[5] Kreuseler, M.; Schumann, H.: A Flexible Approach for Visual Data Mining; IEEE Transactions on Visualization and Computer Graphics, Band 8, Nr.1, Januar-März, 2002, S. 39-51

[6] Kohonen, T.: Self Organizing Maps. Berlin, Springer-Verlag. 1995

[7] John, M.; Tominski, C.; Schumann, H.: Visual and Analytical Extensions for the Table Lens. IS&T/SPIE Annual Symposium Electronic Imaging - Visualization and Data Analysis (VDA), San Jose, USA, 2008.

[8] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda, “Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data,” in Proceedings of IEEE Symposium on Information Visualization (InfoVis'05), IEEE Press, 2005.

[9] dos Santos, S., and Brodlie, K. (2004): Gaining understanding of multivariate and multi-dimensional Data through Visualization. Computers & Graphics, 28: 311- 325, Elsevier.

[10] Wijk, J.J. van and E. van Selow. 1999. Cluster and Calendar-based Visualization of Time Series Data. In: G. Wills, D. Keim (eds.), Proc. IEEE Symposium on Information Visualization (InfoVis'99), IEEE Computer Society, pp 4-9.

[11] Nocke, T.: Visuelles Data Mining und Visualisierungsdesign für die Klimaforschung. Promotion, Universität Rostock, 2007

[12] Müller, W.; Nocke, T.; and Schumann, H.: Enhancing the Visualization Process with Principal Component Analysis to Support the Exploration of Trends. Proceedings Asia Pacific Symposium on Information Visualization (APVIS'06), Tokyo, Japan, Feb. 2006

[13] Johnson, B. and Shneiderman, B.: Treemaps - A Space-Filling Approach to the Visualization of Hierarchical Information Structures. Proceedings of the IEEE Information Visualization '91, S. 275–282, 1991

[14] http://www.vismaster.eu/home