Visual Analytics: Towards Intelligent Interactive Internet and Security Solutions

James Davey[1], Florian Mansmann[2], Jörn Kohlhammer[1], and Daniel Keim[2]

[1] Fraunhofer IGD, Germany
[2] Universität Konstanz, Germany

Abstract. In the Future Internet, *Big Data* can not only be found in the amount of traffic, logs or alerts of the network infrastructure, but also on the content side. While the term Big Data refers to the increase in available data, this implicitly means that we must deal with problems at a larger scale and thus hints at scalability issues in the analysis of such data sets. Visual Analytics is an enabling technology, that offers new ways of extracting information from Big Data through intelligent, interactive internet and security solutions. It derives its effectiveness both from scalable analysis algorithms, that allow processing of large data sets, and from scalable visualizations. These visualizations take advantage of human background knowledge and pattern detection capabilities to find yet unknown patterns, to detect trends and to relate these findings to a holistic view on the problems. Besides discussing the origins of Visual Analytics, this paper presents concrete examples of how the two facets, content and infrastructure, of the Future Internet can benefit from Visual Analytics. In conclusion, it is the confluence of both technologies that will open up new opportunities for businesses, e-governance and the public.

1 Introduction

We live in a world that faces a rapidly increasing amount of data. Today, in virtually every branch of commerce and industry, within administrative and legislative bodies, in scientific organisations and even in private households vast amounts of data are generated. In the last four decades, we have witnessed a steady improvement in data storage technologies as well as improvements in the means for the creation and collection of data. Indeed, the possibilities for the collection of data have increased at a faster rate than our ability to store them [4]. It is little wonder that the buzzword *Big Data* is now omnipresent. In most applications, data in itself has no value. It is the *information* contained in the data which is relevant and valuable.

The *data overload* problem refers to the danger of getting lost in data, which may be: 1. irrelevant for the current task, 2. processed in an inappropriate way, or 3. presented in an inappropriate way. In many application areas success depends on the right information being available at the right time. The acquisition of raw data is no longer a problem: it is the lack of methods and models that can turn data into reliable and comprehensible information.

F. Álvarez et al. (Eds.): FIA 2012, LNCS 7281, pp. 93–104, 2012.

Visual Analytics aims at turning the data overload problem into an opportunity. Its goal is to make the analysis of data transparent for an analytic discourse by combining 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 distinct, complementary capabilities to obtain the best results. The user has the ultimate authority in determining the direction of the analysis. At the same time, the system provides the user with effective means for interaction. Visual Analytics research is interdisciplinary, combining visualisation, data mining, data management, cognition science and other research areas. By fusing the research efforts from these fields, novel and effective analysis tools can be developed to solve the data overload problem.

In this position paper we postulate that Visual Analytics will play a key role in the Future Internet. We consider two facets of the Future Internet: content and infrastructure. Both facets are characterised by vast and growing amounts of data including the following examples: 1. Vast amounts of user-generated content exists in private and public networks. 2. The new trend towards open data means that ever more administrations and NGOs are making their data available online. 3. Simulations of new architectural concepts for the Internet generate vast amounts of data. 4. Huge repositories of network and security data exist and are growing.

Visual Analytics researchers are already developing techniques to address the data overload problem. Thus, we believe that these technologies can make a significant contribution to the success of the Future Internet. With the help of Visual Analytics, the creators and users of the Future Internet will be able to turn data overload from a problem into an opportunity.

The rest of this article is structured as follows: Sect. 2 provides an introduction to Visual Analytics. In the subsequent two sections, an overview of the current and potential uses of Visual Analytics in the Future Internet is presented. In Sect. 3 we focus on content analysis and in Sect. 4 on analysis for the improvement and protection of network infrastructure. We close with a conclusion and outlook in Sect. 5.

2 The Origins of Visual Analytics

Visual analytics emerged as the synthesis of a number of separate disciplines. Most prominent among these were information visualization and data mining. In this section we will briefly introduce each of these fields and then explain how Visual Analytics developed as a new, separate research area.

2.1 Disciplines Contributing to Visual Analytics

Information Visualization (InfoVis) emerged as an independent discipline from the scientific visualization community in the late 1990's. Central to the formalization of the field was the so-called *InfoVis Pipeline* shown in Fig. 1, published in 1999 [3]. In contrast to scientific visualization, InfoVis involved the interactive

visualization of abstract data, i.e. data without an explicit physical or spatial reference. As evidenced by the original InfoVis Pipeline, the first InfoVis techniques were developed for tabular data. Later, techniques were developed or extended to apply to data in other formats, such as data cubes, graphs and text collections.

The goal of information visualization is to use images derived from data as a means to assist users in their exploration of large data sets. Thus, it aims to allow people to use their strongest sense, *vision* to think [3]. The late 1990's and the first years of this century saw an explosion in the number and diversity of published visualization techniques. In the last five years, the research focus of the InfoVis community has shifted to the evaluation of these techniques and the development of best practices.

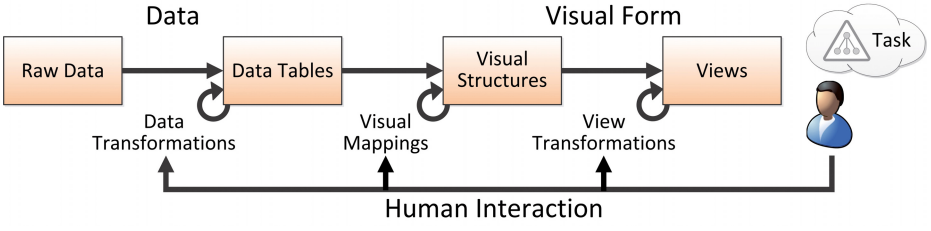

Fig. 1. The InfoVis Pipeline; based on the pipeline presented in [3]

Data Mining was also born in the 1990's from the need to explore and analyse large amounts of data. The field was first formalised in the book *Knowledge Discovery in Databases* (KDD) in 1991 [17]. The so-called *KDD pipeline* shown in Fig. 2 was defined in a subsequent book in 1996 [5].

In a broad sense, data mining involves the use of statistical and machine-learning techniques to discover *patterns* in large data sets. Data mining tasks include the characterization or description of data subsets, the mining of rules describing associations or correlations, classification or regression for predictive analysis, cluster analysis and outlier analysis [9]. Initially, these techniques were focused on relational database management systems. However, the field has developed to include techniques for the analysis of a great variety of data sources, including text collections, video, image and spatio-temporal data.

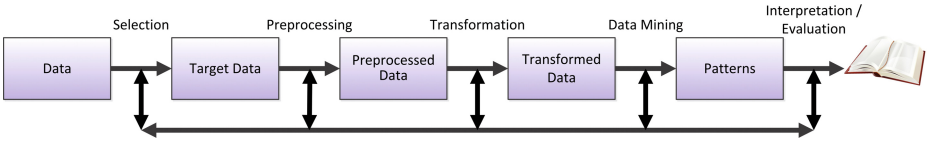

Fig. 2. The KDD Pipeline; based on the pipeline presented in [5]

2.2 Definition of Visual Analytics

Visual Analytics was first defined by Thomas et al. as "the science of analytical reasoning facilitated by visual interactive interfaces" [19]. It emerged as an attempt to compensate for the deficits of both data mining and information visualization. The results of data mining algorithms are frequently difficult to understand and often even more difficult to share with others. This lack of transparency demanded a means to *see* the models, parameters and assumptions on which those results were based.

Information Visualization provides techniques which allow human users to examine abstract data with the help of visualizations. These can also be used to expose the details of automated analysis steps. The Visual Analytics process as proposed by Keim et al. is shown in Fig. 3 [12].

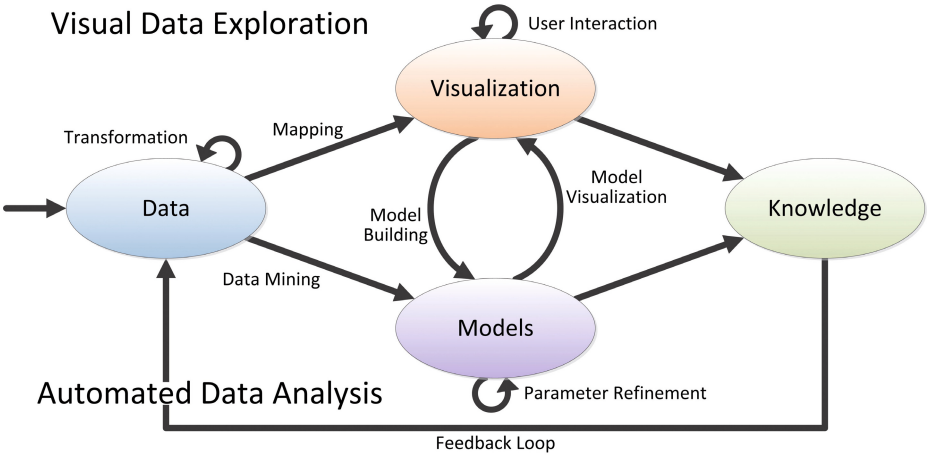

Fig. 3. The Visual Analytics Process; first presented in [12]

In 2009 and 2010 a Coordinated Action named *VisMaster* and funded by the European Commission set out to establish a Visual Analytics research community in Europe. The primary result of the project was a research roadmap entitled *Mastering the Information Age* [11]. The established community has continued its work after the project. Its main channel for dissemination and coordination of community activities is the European Visual Analytics website.[1]

3 Visual Analytics for Internet Content

Despite the fact that Visual Analytics developed largely independently of Future Internet technologies, the current trend towards visualization toolkits for the web

[1] `http://www.visual-analytics.eu`

(e.g., D3 [2] or Polymaps[2]) suggests that visualization will play a greater role in the near future. In addition, the number of organisations publishing their data online is growing. As a result, new opportunities for linking and exploring these open data sets in the Future Internet with the help of Visual Analytics arise.

3.1 Open and Public Data

During the last decade the visualization community began the creation of interactive web visualizations that empowered the public to investigate open data sets on their own. One of the most successful approaches was IBM's ManyEyes platform [20], which allows users to upload data, visualize it either statically or interactively and then facilitates discussions about findings within the user community. Besides well known charts, such as scatter plots, bar, line and pie charts, the platform features more advanced visualizations, such as tree maps, stacked graphs and bubble charts, and a number of textual visualizations, such as word clouds, phrase nets and word trees. Newer web visualization tools, such as Google's Public Data Explorer[3] and Tableau Public[4] extend both the accessibility of data as well as the diversity of available web visualization tools.

While web visualization tools for open data have already started to emerge, the combination of visualization and data mining tools in Visual Analytics applications are not yet available for the web. However, we expect them to emerge in a new wave of Visual Analytics frameworks and tools for the web.

3.2 Smart Cities

Smart Cities are characterized by their competitiveness in the areas smart economy, smart mobility, smart environment, smart people, smart living and smart governance [8]. While strengths in each of these areas have strong links to the historic development of cities, technological advancements such as the Future Internet or Visual Analytics can play a role in boosting their competitiveness.

As an example, Visual Analytics applications such as the one detailed in the study [1] can significantly empower the analysis of traffic conditions (e.g. traffic jams) using data from GPS tags of a sample of the total vehicle population within the city. Future Internet technologies not only play a role in the data collection infrastructure (Internet of Things), but also in the propagation of analysis results to commuting citizens. However, Visual Analytics is required to turn the large and complex mobility data into useful information.

Smart governance can be enhanced through the combination of Visual Analytics and Future Internet technologies by analysing available data in the detailed geographic context of the city. MacEachren et al. [15], for example, created a Visual Analytics tool that takes advantage of a geo-tagged Twitter stream for the assessment of situational awareness in application scenarios ranging from

[2] http://polymaps.org/
[3] http://www.google.com/publicdata/
[4] http://www.tableausoftware.com/public

disease monitoring, through regional planning, to political campaigning. As demonstrated in this and the previous examples, it is the combination of Visual Analytics and Future Internet technologies that enables the advancement of opportunities for Smart Cities.

We believe that all areas of Smart Cities can benefit from Visual Analytics and Future Internet technologies to maintain and increase their attractiveness for their citizens, companies and institutions.

3.3 Text and News Analysis

The Internet is full of unstructured, but often interlinked, data that could potentially be valuable if processed and presented in a meaningful way. However, issues of data processing (e.g., data quality or entity recognition) and representation (e.g., usability or scalability) turn such efforts into challenging undertakings and only very focused approaches have so far succeeded.

Fig. 4, for example, shows a Visual Analytics system for the analysis of online news [14] collected by the *Europe Media Monitor*[5]. Text clustering is used to extract stories from the news articles and to detect merging and splitting events in such stories. Special care is taken to minimize clutter and overlap from edge crossing while allowing for incremental updates. Besides the main entity and daily keywords for each story, the figure shows a list of emerging stories at the top and a list of disappearing stories at the bottom of the screen.

Text mining can be useful for the automatic extraction of opinions or sentiments from user-generated content. While this data in itself is valuable, making sense of a large collection of results can be supported using visualization as demonstrated in the study of Kisilevich et al. [13] dealing with photo comments.

In summary, the use of Visual Analytics in the Future Internet for the analysis of text and news data can lead to innovative web applications. However, the unstructured nature and the linguistic intricacies of processing large but possibly short (e.g. Twitter postings) textual data generated by a multitude of people in several languages can impose significant challenges on the processing side.

3.4 Future Work

Currently, three projects funded by the European Commission are addressing the challenges of Smart Governance. The projects will make use of opinion mining and visualization technologies to draw on user-generated Internet content to inform policy-making decisions. The *ePolicy* project[6] is focused on the policy-making life cycle in regional planning activities. The life cycle integrates global concerns (e.g. impacts, budget constraints and objectives) and individual perspectives (i.e. opinions, reactions extracted from the web) into the decision process, giving guidance towards better policy implementation strategies.

[5] `http://emm.newsbrief.eu/`
[6] `http://www.epolicy-project.eu`

Fig. 4. Visual Analytics for news story development. Stories are extracted from online news articles and visualized over several days. Distinct stories about "Omar Suleiman" and "Tahrir Square" partly merge on the 9th of February. On the 10th of February a linked story involving the "White House" emerges.

The *NOMAD* project[7] aims to provide politicians with tools to draw on non-moderated sources in the Social Web for the policy appraisal. A focus will be laid on the presentation of arguments drawn from relevant constituencies for and against policy decisions. The *FUPOL* project[8] aims to combine simulations of the effects of policy decisions with information drawn from the Social Web, as well as crowd-sourcing techniques. FUPOL will target domains such as sustainable development, urban planning, land use, urban segregation and migration.

While most of the interactive Visual Analytics applications currently run as stand-alone applications, we believe that in the near future these applications will not only take advantage of the open and public data available in the web, but move towards client-based applications running in modern web browsers. Furthermore, we are convinced that data linkage, text mining and modern data management approaches will open up new opportunities for the inclusion of Visual Analytics in Future Internet technologies. This is further supported by the fact that streaming text data visualization (cf. [18]) is currently a hot topic in the visualization and Visual Analytics research community.

[7] `http://www.nomad-project.eu`
[8] `http://www.fupol.eu`

4 Visual Analytics for Network Infrastructure

The growing complexity of network infrastructure cries out for more analytical support on both the automated side as well as on the human side. While we have witnessed an exponential growth in networking and computing capacities, the number of persons involved in maintaining our networks has not expanded in the same way. Our only chance to tackle the networking issues of the Future Internet are to either manage tasks automatically or to empower network administrators to tackle large scale issues in a more efficient way. This section will discuss how we can combine automated and visual approaches through Visual Analytics to keep the Future Internet's infrastructure alive.

4.1 Infrastructure Planning and Testing

Network infrastructure planning and testing are complex tasks. Besides historic capacity utilization statistics, forecasting plays an important role. However, since comprehensive interpretation of the huge volumes of data exceeds human capacities, meaningful abstractions and a focus on specific sub-problems are necessary to master the network's complexity. Visual Analytics builds on human perceptual capabilities to spot interesting patterns and automatic methods to deal with large scale data and thus enables interpretation at a higher level of detail. Furthermore, interaction methods extend Visual Analytics methods and enable exploratory analysis tasks.

Hierarchical Network Maps [16] are one example of how visualization can facilitate the interpretation of network capacities. In particular, this technique uses a hierarchy of continents, countries, autonomous systems and IP prefixes to render a TreeMap [10] of the internet. Coloring can then be used to match traffic load onto rectangles and interaction facilitates drill-down along the levels of the hierarchy for chosen regions.

4.2 Network Security

Today, signature-based and anomaly-based intrusion detection are considered state-of-the-art in network security. However, fine-tuning parameters and analysing the output of these intrusion detection methods can be complex, tedious, and even impossible when done manually. In general, systems become more and more sophisticated and make decisions on their own up to a certain degree. However, as soon as unforeseen events occur, system administrators or security experts have to intervene to handle the situation. While network monitoring and security have profited a lot from automatic detection methods in recent years, visual approaches foster a better understanding of the complex information through interactive visualization and therefore have a lot of potential to complement the former approaches.

By means of the Visual Analytics application NFlowVis [7] we demonstrate in this section how the combination of automatic and visual analysis can help security experts to derive more meaning out of the vast amount of security events

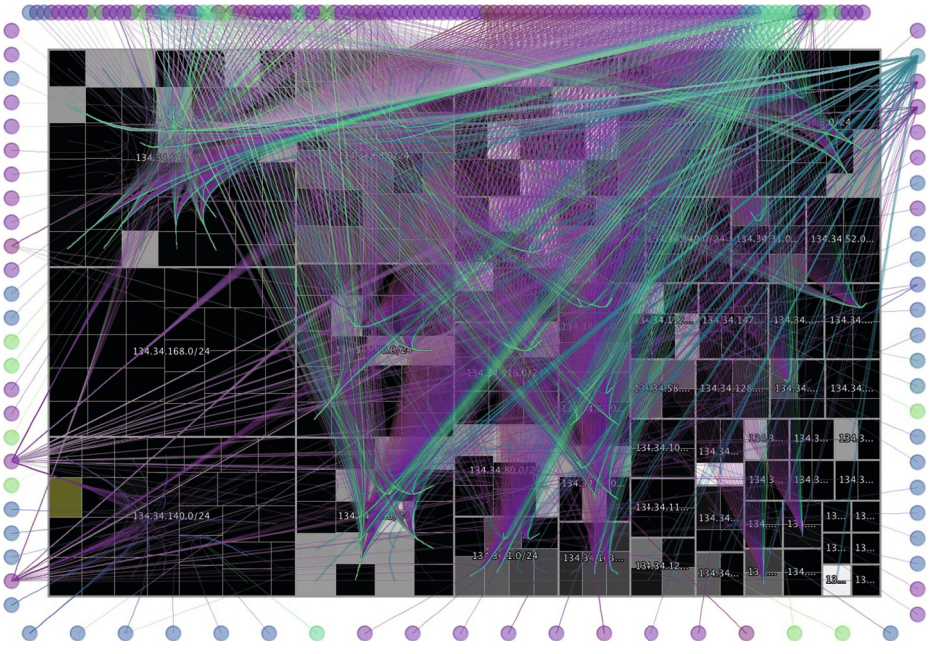

Fig. 5. Analysis of a distributed network attack on the SSH service of a university network on May 11, 2008 using NFlowVis [7]. The circles on the outside represent hosts that attack the squared hosts in the internal network. A clustering algorithm ensures that related attackers (see top) are positioned next to each other.

and traffic data which is characteristic of the field. In particular, we use traffic patterns which are common for signature-based intrusion detection systems and one day of network traffic statistics (NetFlows) from the main gateway of a medium sized university, which amounts to approximately 10 GB of raw data.

Fig. 5 shows the visual output of an analysis with NFlowVis. After having selected suspicious hosts from the intrusion detection system, their network traffic to all hosts in the internal network is retrieved from a database and visualized. While automatic intrusion detection systems output many alerts in a large network, the visualization supports the analyst in the difficult task of correlating these alerts with each other and setting them into context. In this particular case, we chose an SSH traffic pattern and visualized a number of external hosts matching this traffic pattern.

Before visualizing the information, the system first clusters the external hosts (potential attackers) and then places them on the nearest border in such a way that a) hosts with similar traffic patterns appear next to each other and b) preferably short splines are drawn to connect the dots of the external hosts and the rectangles representing their internal communication partners. Color encodes the first byte of the IP address of the external host in such a way that attackers from nearby network prefixes are drawn in a similar color. This helps

to judge whether the attack is conducted from a particular network or from hosts distributed all over the Internet.

Drawing straight connecting lines results in a lot of visual clutter. To reduce this clutter, the lines are grouped by exploiting the structure of the underlying hierarchical visualization of the /24 prefixes. As a result, the analyst can easily identify the pattern of the distributed attack on the upper right of Fig. 5, which details a number of external hosts targeting the same subset of internal hosts in the university network. A more detailed analysis revealed that all attacking hosts contacted 47 hosts and thereby consciously avoided a common threshold of an automatic intrusion detection system. The visual output furthermore shows scanning activity of individual hosts on the lower left and top right of Fig. 5. We assume that scanning activity first identified candidate victims in the network and that the botnet then used this information to target this subset of hosts in the university network, since the number of attacked hosts per subnet varies.

Currently, the *VIS-SENSE* project[9], funded by the European Commission, is applying Visual Analytics techniques to large, network-security-related data sets. The project focuses on the strategic analyses of spam, malware and malicious websites. In addition, the misuse of the Border Gateway Protocol for criminal activities will be analysed.

4.3 Real-Time Monitoring

Modern services heavily rely on the availability of the network and server infrastructure to comply with the strict service level agreements of business users and consumers. However, defining a valid state for all components of the network is not possible due to the high number of complexities and inter-dependencies of all involved systems. Modern monitoring approaches therefore often produce either too many or too few alerts, which makes manual analysis close to real-time almost impossible.

In this case Visual Analytics can bridge the gap between the complexity of the data and the human understanding and thus speed-up both investigation of failures and system recovery operations. The work in [6], for example, details a Visual Analytics system for the analysis of system log events in real-time. With peaks of up to 425,000 events per hour, the interactive time-line visualization and the geographic map interface highlight events according to a scoring model and enable the detection of unusual activity, such as remote accesses from uncommon sources or bursts of critical events on servers.

4.4 Future Work

While this section detailed some exemplary uses of Visual Analytics for planning, monitoring and securing network infrastructure, many tasks in this wide field are still conducted without any visual or computational support. We therefore see a lot of potential for research that connects the still largely independent fields of Visual Analytics and the Future Internet.

[9] `http://www.vis-sense.eu`

5 Conclusion

In this article we presented an introduction to Visual Analytics and its relevance for the Future Internet. We considered the two facets content and infrastructure. Both facets are characterized by a vast and growing amount of data.

With respect to content in the Future Internet, we have shown that emerging data visualization platforms for the web derive their value from the relevance of the data that is analysed with them. Since more and more open and public data becomes available every day, it is only a matter of time before existing visualization platforms hit scalability limits – due to the data overload problems at hand – and need to include automated data analysis functionality. While the analysis of the abundance of text and news available in modern media like Twitter imposes significant challenges, working on these problems can have drastic effects on the development of countries, regions and smart cities. We are thus convinced that targeted research in Visual Analytics can revolutionize the way in which we interact with content in the Future Internet.

Besides its potential for content, Visual Analytics can play an important role in the network infrastructure of the Future Internet. Due to the amount of data available from networking devices, the inherent complexity of the network and the need to immediately react to failures or attacks, visual and computational support for tasks in this domain can significantly improve infrastructure planning and testing, as well as network monitoring and security. We conclude that strengthening the connection between Visual Analytics and the Future Internet will enable us to build a more secure, reliable and scalable network.

The examples presented show how Visual Analytics is already contributing solutions to the data overload problem in the Future Internet. Thus we are convinced that the confluence of both technologies has enormous potential for use in the business, administrative and private spheres.

References

1. Andrienko, G.L., Andrienko, N.V., Hurter, C., Rinzivillo, S., Wrobel, S.: From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In: IEEE Conference on Visual Analytics Science and Technology (VAST 2011), pp. 161–170 (2011)
2. Bostock, M., Ogievetsky, V., Heer, J.: D3: Data-driven documents. IEEE Transactions on Visualization and Computer Graphics 17(12), 2301–2309 (2011)
3. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Readings in information visualization: using vision to think. Morgan Kaufmann Publishers Inc., San Francisco (1999)
4. Cukier, K.: Data, data everywhere: A special report on managing information. The Economist 1(1), 14 (2010)

5. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI Press (1996)
6. Fischer, F., Mansmann, F., Keim, D.A.: Real-Time Visual Analytics for Event Data Streams. In: Proceedings of the 2012 ACM Symposium on Applied Computing, SAC 2012. ACM (2012)
7. Fischer, F., Mansmann, F., Keim, D.A., Pietzko, S., Waldvogel, M.: Large-Scale Network Monitoring for Visual Analysis of Attacks. In: Goodall, J.R., Conti, G., Ma, K.-L. (eds.) VizSec 2008. LNCS, vol. 5210, pp. 111–118. Springer, Heidelberg (2008)
8. Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., Meijers, E.: Smart cities ranking of european medium-sized cities (2009), http://www.smart-cities.eu/ (retrieved January 20, 2012)
9. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques., 3rd edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers Inc., Waltham (2012)
10. Johnson, B., Shneiderman, B.: Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In: Proc. IEEE Conference on Visualization, pp. 284–291. IEEE (1991)
11. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F. (eds.): Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics (2010), http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf
12. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual Analytics: Scope and Challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
13. Kisilevich, S., Rohrdantz, C., Keim, D.A.: Beautiful picture of an ugly place. Exploring photo collections using opinion and sentiment analysis of user comments. In: Computational Linguistics & Applications (CLA 2010), pp. 419–428 (October 2010)
14. Krstajic, M., Najm-Araghi, M., Mansmann, F., Keim, D.: Incremental Visual Text Analytics of News Story Development. In: Proceedings of Conference on Visualization and Data Analysis, VDA 2012 (2012)
15. MacEachren, A.M., Jaiswal, A.R., Robinson, A.C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., Blanford, J.: Senseplace2: Geotwitter analytics support for situational awareness. In: IEEE Conference on Visual Analytics Science and Technology (VAST 2011), pp. 181–190 (2011)
16. Mansmann, F., Keim, D.A., North, S.C., Rexroad, B., Sheleheda, D.: Visual Analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats. IEEE Transactions on Visualization and Computer Graphics 13(6) (2007)
17. Piatetsky-Shapiro, G., Frawley, W.J. (eds.): Knowledge Discovery in Databases. MIT Press (1991)
18. Rohrdantz, C., Oelke, D., Krstajic, M., Fischer, F.: Real-Time Visualization of Streaming Text Data: Tasks and Challenges. In: Workshop on Interactive Visual Text Analytics for Decision-Making at the IEEE VisWeek 2011 (2011)
19. Thomas, J.J., Cook, K.A. (eds.): Illuminating the Path: the Research and Development Agenda for Visual Analytics. IEEE CS Press (2005)
20. Viegas, F., Wattenberg, M., Van Ham, F., Kriss, J., McKeon, M.: Manyeyes: a site for visualization at internet scale. IEEE Transactions on Visualization and Computer Graphics 13(6), 1121–1128 (2007)