Solving Problems with Visual Analytics: Challenges and Applications

Daniel A. Keim, Leishi Zhang, Miloš Krstajić, Svenja Simon
University of Konstanz
Germany
leishi.zhang@uni-konstanz.de

ABSTRACT**:** *Never before in history data has been generated and collected in such high volumes as it is today. Keeping up to date with the flood of data, using standard tools for data analysis and exploration, is fraught with difficulty. Visual analytics seeks to provide people with better and more effective ways to understand and analyze large datasets, while also enabling them to act upon their findings immediately. The field integrates the analytic capabilities of the computer and the abilities of the human analyst, allowing novel discoveries and empowering individuals to take control of the analytical process. In this paper we present the challenges of visual analytics and exemplify them with several application examples that illustrate the existing potential of current visual analysis techniques but also their limitations.*

Keywords: Spatial-temporal Data Analysis, News Stream Data Analysis, Next-Generation-Sequencing Data Analysis, Visual Readability Analysis

Received: 12 December 2011, Revised 1 February 2012, Accepted 7 February 2012

1. Introduction

One of the greatest challenges in the 21st century is to handle massive data collections. The fast development of data storage devices and means to collect data enables people in science and business domains to gather vast amounts of information from different sources. For instance, by 2003 information on 20,000-25,000 genes in the human DNA and the sequences of the 3,000,000,000 chemical base pairs that make up the human DNA had been collected by the Human Genome Project to help understanding how the human body functions [3]. In a different application, every day about 100,000 articles from over 2,500 news sources in 43 languages are collected by the Europe Media Monitor (EMM) to aggregate worldwide news and issue alerts [2]. These data are rich information sources that can help to support scientific discovery and decision making. However, extracting meaningful knowledge from such data is challenging. People are often confronted by disparate, conflicting, and dynamic information from multiple heterogeneous sources and get lost due to the lack of ability to analyze it. This is the well-known *information overload problem*.

As a young upspring science field that aims at tackling the information overload problem, visual analytics *combines automated data analysis with interactive visualizations for an effective understanding, reasoning, and decision making on the basis of very large and complex datasets* [5]. The field utilizes both the computational power of computers and the innate human ability to visually perceive patterns and trends to help people extract meaningful patterns from data that are too large or too complex to be handled by automated data analysis methods or visualization alone. The essential idea is to develop computer based intelligent systems that allow human analysts to examine the massive information stream at the right level of abstraction through appropriate visual representations and to take effective actions in real-time.

Figure 1 illustrates the process of visual analytics. First of all, heterogeneous data sources need to be processed and integrated. Automated analysis techniques can then be applied to generate models of the original data. These models can be visualized for evaluation and refinement. In addition to checking the models, visual representations can be abstracted from the data using a variety of interactive visualization techniques that are best suited for the specific data type, structure, and dimensionality. In the visual analytics process, knowledge can be gained from visualization, automatic analysis, as well as the interactions between visualization, models and the human analysts. The feedback loop stores this knowledge of insightful analysis in the system and assists the analyst in drawing faster and better conclusions in the future. An important aspect of visual analytics is user interaction: analysts should be able to query and explore, often multiple abstract views of the data, with rapid feedback, and in the same way steer the analysis process by modifying parameters or choosing alternative analysis methods.

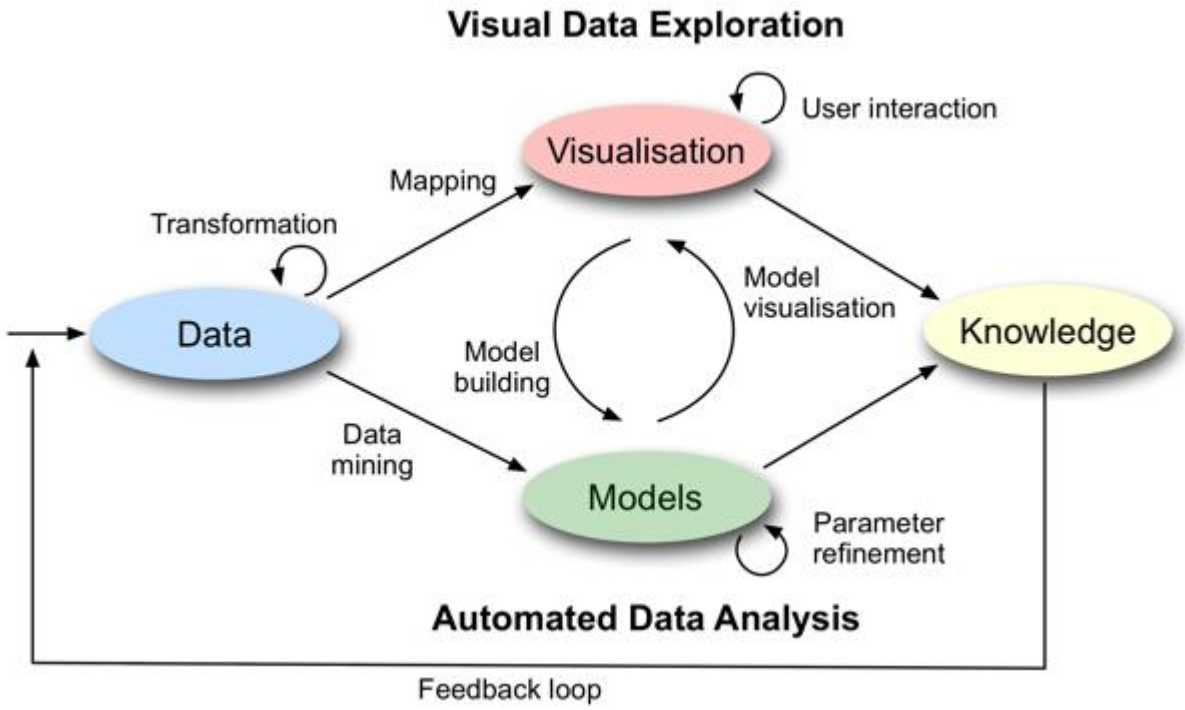

Figure 1. Visual Analytics Process [5]

Visual analytics is essential in many application areas where large information spaces have to be processed and analyzed. However, challenges exist in many aspects of the field alongside the opportunities. In this paper, we discuss these challenges and exemplify them with some real world applications. In section 2, we briefly introduce related work in defining the concept, scope and paradigm of visual analytics. In section 3, we detail the challenges of visual analytics in the technical aspects of the field. In section 4, we use some example applications to illustrate the potential of current visual analytics techniques, and in section 5 we provide some concluding remarks.

2. Related Work

The field of intelligent data analysis dates back to automated data analysis, which makes use of the computational intelligence of modern computers to extract patterns from data using statistics or data mining methods. More than a decade ago the information visualization community recognized the potential of improving the effectiveness of automated data analysis by providing visual aids and getting users more involved in the knowledge discovery process [15]. The integration of information visualization and automated data analysis lead to the field of visual data mining, which widened the scope of both fields and resulted in new techniques and many interesting and important research opportunities. Visual analytics evolved from visual data mining by bringing more human analytical skills into the loop and further exploiting human's perceptive and cognitive capabilities. Through interactive user interfaces, the analysts can take control of the analytical process such that the direction of the data exploration can be decided based on real time findings. This is well suited for exploratory analysis of large complex data, especially when the goals are not clearly defined.

The main objective of visual analytics to present, navigate, aggregate, and see the details of the data such that complex questions can be answered. Shneiderman [13] proposed the well-known information seeking paradigm '*Overview First, Zoom and Filter, Details on Demand*'. The '*golden rule*' starts with an overview for a better orientation, lets the user decide how to filter interesting data, and shows details only on demand for selected data. It has successfully guided the design of many visualization systems and techniques. For visual analytics, we extend Shneiderman's paradigm to: '*Analyze First; Show the Important; Zoom, Filter and Analyze Further; Details on Demand*' [4]. The paradigm brings more analytical skills into the loop before and after the generation of interactive visual representations to help extract abstract models from data sets that are too large or too complex to be analyzed in a straightforward manner.

The core of visual analytics is visualization. Visualization acts as a communication platform of the other building blocks of visual analytics that come from closely related disciplines including data management, data mining, spatialtemporal data analysis and human perception/cognition. Visual analytics also requires an appropriate infrastructure in terms of software and data sets and related analytical problems repositories, as well as evaluation facilities to assess solutions across all disciplines (see Figure 2). Details of these building blocks and their relationships can be found in [5].

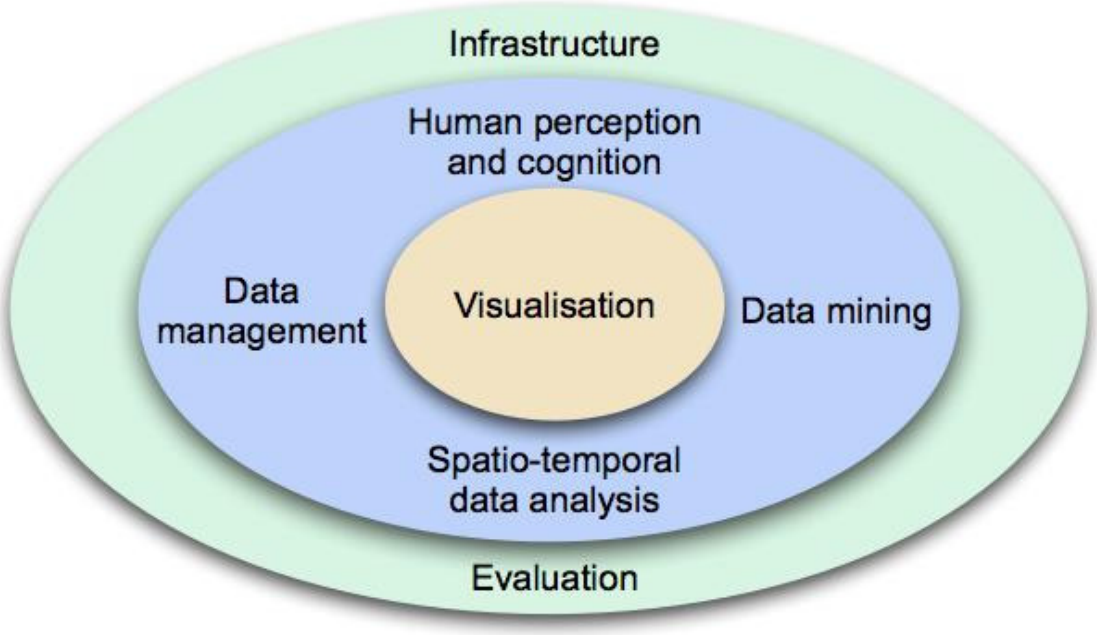

Figure 2. Building Blocks of Visual Analytics [5]

3. Challenges

Visual analytics is a growing field with many open problem and challenges. Many challenges originate from the specific applications of visual analytics. Each individual application has its own practical requirements in its particular problem domain. However, some challenges are common to more than one domain and application:

3.1 Scalability

It is difficult to come up with scalable visual analytics solutions with regard to both visual representations and automatic analysis. The solution needs to scale in size, dimensionality, data types, and levels of quality. Effective methods are needed to deal with noisy high-resolution input data as well as continuous input data streams of high bandwidth. The relevant data patterns and relationships need to be visualized on different levels of details, and with appropriate levels of data and visual abstraction.

3.2 Uncertainty

Dealing with uncertainty in visual analytics is nontrivial because of the large amount of noise and missing values originating from heterogeneous data sources and bias introduced by automatic analysis methods as well as human perception.

To face this problem, the notion of data quality and the confidence of the analysis algorithm need to be appropriately represented. The analysts need to be aware of the uncertainty and be able to analyze quality properties at any stage of the data analysis process.

3.3 Interestingness

To extract meaningful patterns and highlight potential events in the data, appropriate measures need to be defined such that the '*nteresting*' proportions of data can be identified easily. The interestingness of a data fragment can vary a lot depending on the nature of data and user needs. This implies that features with general validity need to be defined. In addition, the interestingness measures have to be defined at a generic level such that different types of data (numerical, text, etc.) can be handled. Furthermore, effective algorithms need to be designed to enable the user to extract meaningful patterns from the data using the interestingness measures.

3.4 Text Data Stream

Text data (e.g. news, Twitter posts) often provide useful information for understanding the actual situation in case of a sudden event. While event-changes based time series data analysis is a well-studied topic in the context of numerical data [6], analysis and visualization of text steams is still a relatively new field with many open issues. Often, text stream data have little structure, grammar and context and are provided as high-frequency multilingual stream containing a high percentage of non-meaningful and irrelevant messages. The challenge is to handle the dynamic nature of the stream data and the low-tolerance of monitoring delay in emergency cases.

3.5 Hardware

More efficient computational methods and powerful hardware are needed to support near real time data processing and visualization for large data streams. In addition to high-resolution desktop displays, advanced display devices such as large-scale power walls and small portable personal assistants need to be supported. Visual analytics systems should adapt to the characteristics of the available output devices, supporting the visual analytics work-flow on all levels of operation.

3.6 Interaction

Novel interaction techniques and tangible user-interfaces are needed to fully support the seamless intuitive visual communication with the system. The analyst should be able to fully focus on the task at hand and not be distracted by overly technical or complex user interfaces and interactions. User feedback should be taken as intelligently as possible, requiring as little user input as possible. Such interactions guarantee the full support of the user in navigating and analyzing the data, memorizing insights and making informed decisions.

3.7 Evaluation

Due to the interdisciplinary nature and complex visual analytics process, it is hard to assess the quality of visual analytics solutions. A theoretically founded evaluation framework needs to be developed to assess the effectiveness, efficiency and user acceptance of new visual analytics techniques, methods, and models. Such a framework will lead to a better understanding of the field and more successful and efficient development of innovative methods and techniques.

3.8 Infrastructure

So far, most current visual analytics solutions develop their own infrastructures for solving their specific problems. Although some systems can be connected through various communication mechanisms such as direct library linking and web services, there is still a mismatch between the level of service provided and the real need for visual analytics in terms of fast and precise answers with progressive refinement, incremental re-computation, and steering the computation towards data regions that are of higher interest to the user. More research is needed to develop a high level infrastructure to bind together all the processes, functions, and services supplied by various disciplines. There is also a need to build repositories of available visual analytics solutions to ensure that common components are reusable.

4. Applications

Visual analytics is a highly application oriented discipline driven by practical requirements. In this section, we use some visual analytics applications to exemplify the possibilities of combining human analytical skills and computer intelligence to solve complex problems. These applications relate to some of the challenges we highlighted in the previous section, for example, pixel based visualization for large scale spatial-temporal data analysis (Spatial-temporal Data Analysis Application), analysis and

visualization of text stream data (News Stream Data Analysis Application), uncertainty handling and interestingness definition in biological applications (Next-Generation-Sequencing Data Analysis Application), and combining human knowledge with automatic analysis via interactive data analysis and visualization (Visual Readability Analysis Application).

4.1 Spatial-temporal Data Analysis

Spatial-temporal data analysis works on data with contain both space and time dimensions, for example, telecommunication

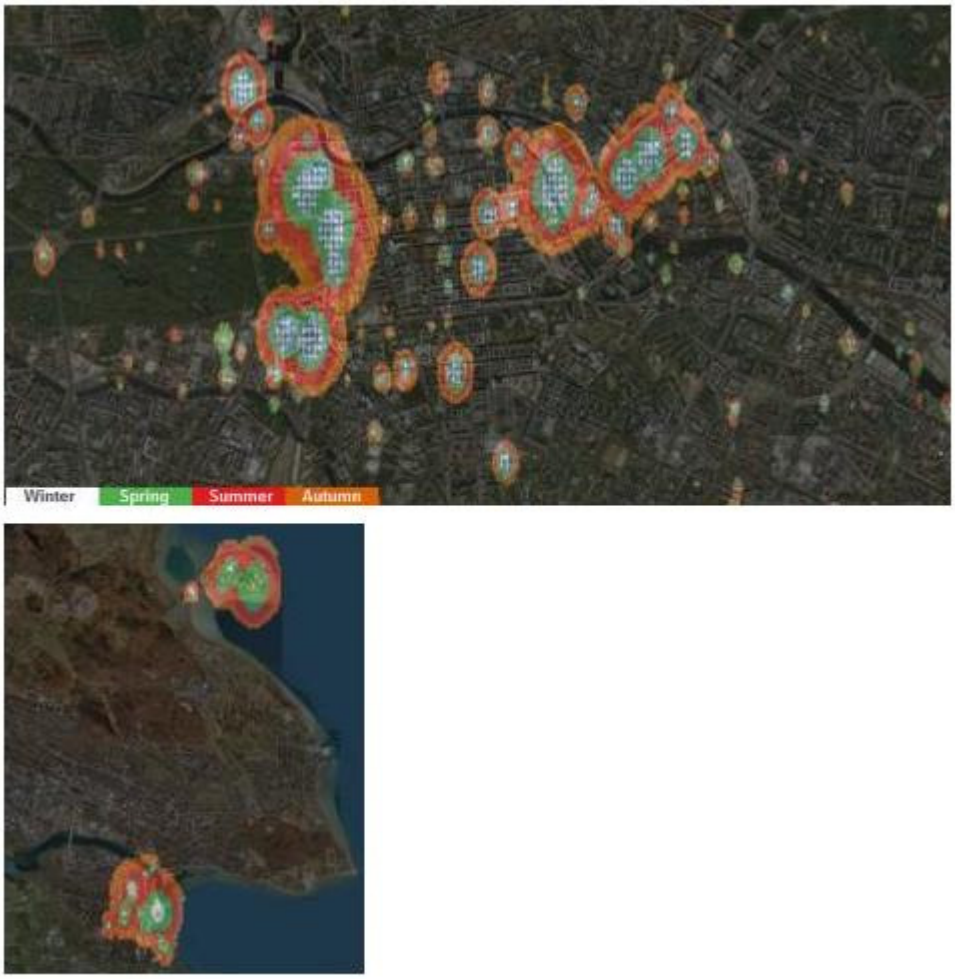

Figure 3. Location and seasonal differences of photos taken in Berlin (top) and Konstanz (bottom): more places of interest are identified in Berlin, and the small size of the white region in Konstanz indicates that this region is mostly visited during the warm seasons, whereas Berlin is visited all year round [1]

traffic data or satellite image data. Spatial data contain measurements together with links to geographic location. Finding spatial relationships and patterns within the data is of particular interest, requiring the development of appropriate management, representation and analysis functions. Temporal data on the other hand records measurements taken over time. Important tasks here include identification of patterns, trends, and correlations of the data items over time. The analysis of data with both spatial and temporal dimensions adds complexities of scale and uncertainty.

Our first application analyzes concentrations and movements of tourists in a city, based on GPS-referenced photos taken at different geographical locations and published online by the tourists. Such data are available from free online photo album services such as Flickr (www.flickr.com) or Panoramio (www.panoramio.com). The Growth Ring Map [1] is a newly developed technique that supports the exploration of both the frequencies and temporal patterns of events occurring at close-by places. The method defines significant places based on the density of a spatial clustering of the photos. To visualize all photos taken

at one significant place, all photos taken at the place are first sorted by their time stamps and then projected on top of the map. Each photo is represented as a colored pixel around the central point (where the photo was taken) in an orbital layout, the older the photo, the closer it is to the central point. Colorhue is used to map semantic properties of the photo, time or place. For example, seasonal differences in visiting the places may be investigated by mapping the seasons to four distinct colors (winter-white, spring-green, summer-red, and autumn-blue). The resulting map shows both the intensity of photos taken at different locations and the seasonal differences (see Figure 3 for an example).

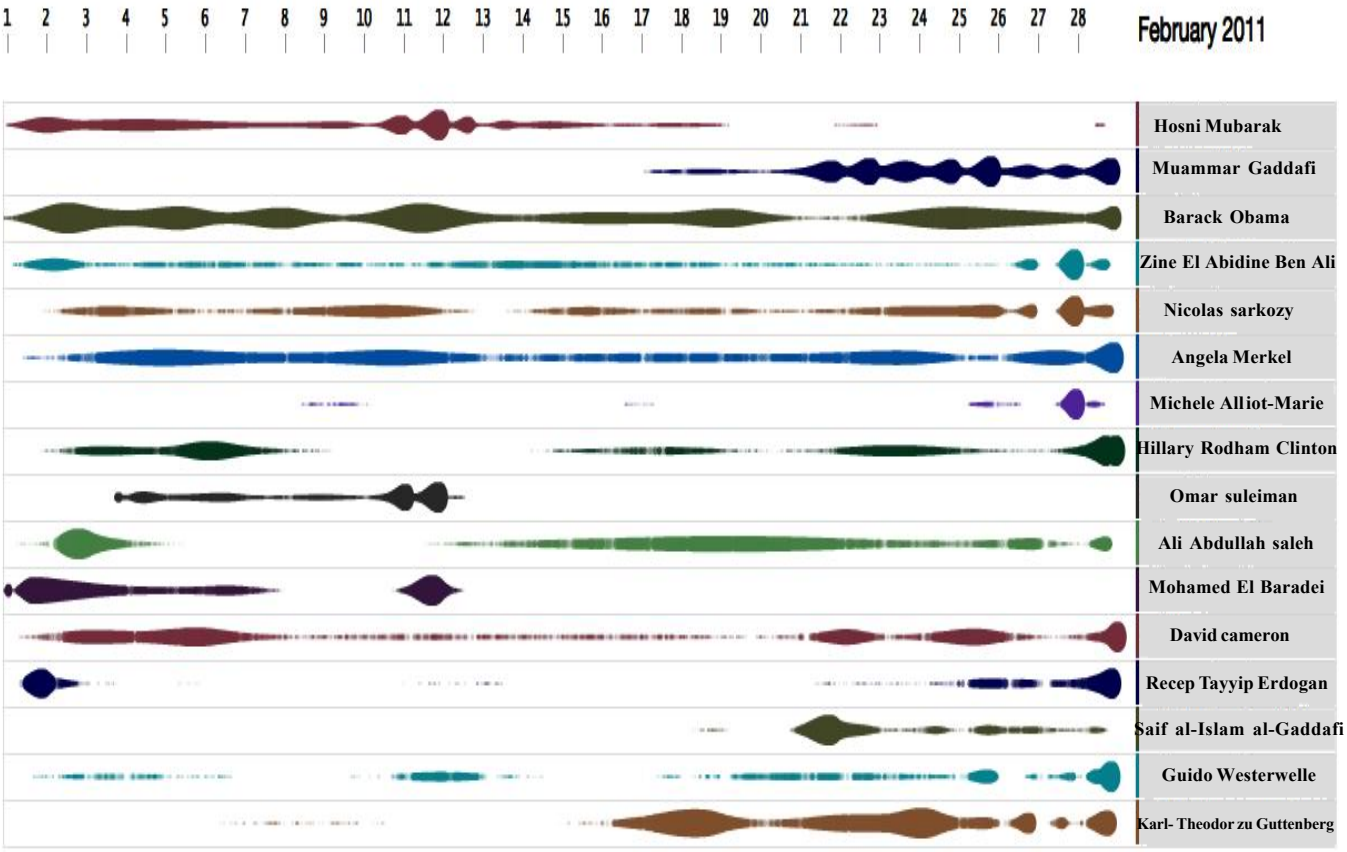

Figure 4. Cloudline: Compact Display of Event Episodes in Multiple Time-Series. The visualization shows multiple politicians appearing frequently in the news during February 2011

4.2 News Stream Data Analysis

The analysis of large quantities of news is an interesting application of visual analytics. A large number of news sources publish thousands of news articles on world events every day and a large number of qualitative and quantitative measures can be calculated from each news article. Making sense of this data is becoming increasingly complex due to the rate of the incoming news, as well as the inherent complexity of analyzing large quantities of evolving unstructured text corpora. Therefore, automated text analysis techniques have to be coupled with user interaction and human knowledge feedback.

Analyzing important people and organizations and their relations to world events is of great interest to users in different application areas. Tracking the appearance of key public figures in the news creates a new and rich information space, in which the temporal context plays an important role. In such cases, it is often necessary to provide an overview of the temporal development of parallel news stories, while allowing the analyst to focus on each individual news article that is a part of an important story. Our tool employs the novel Cloudline [8] technique, which efficiently utilizes limited screen space to present multiple time series within a focus-plus-context display. In Figure 4, each row represents a politician appearing frequently in the news during February 2011. Each news article is represented by a circle, whose size and opacity is determined by its local temporal density. By using kernel density estimation methods on the news time-stamped data, important clusters with high density can easily be identified. The benefits of the method of bypassing time series aggregation are two-fold: first, fine-grained

structure of each time-series is preserved and, second, direct access to the individual items is possible. The tool allows the user to logarithmically distort the timeline and put emphasis on the more recent events. Using magnification lenses with different parameters, a time window of any range can be explored in detail, providing atomic access to each individual news article even within a logarithmically distorted time line. The tool provides flexibility in terms of adapting the technique to different amounts of data, time ranges and tasks. By combining interaction and visual mapping of automatically processed text content, the user is supported by the system in analyzing the evolving corpora and putting each news article into the context of global news trends.

4.3 Next-Generation-Sequencing Data Analysis

The Next-Generation-Sequencing (NGS) technologies made it possible to sequence large amounts of DNA sequences in a short time period and for lower costs, thus opening numerous new possibilities for genomic research, such as determination of structural variations and identification of single nucleotide polymorphisms. By sequencing indirectly RNA (RNA-seq), NGS can also generate data for transcribed genes, which allows many analysis tasks such as gene expression analysis and identification of new genes.

A genome carries the genetic information of living organisms. It is encoded in DNA (or RNA for many viruses) which can be represented as a double-stranded sequence made up of four chemical building blocks (called bases, abbreviated as A, T, C, and G). The informational units in genomes that carry information are the genes.They encode proteins which are functional and

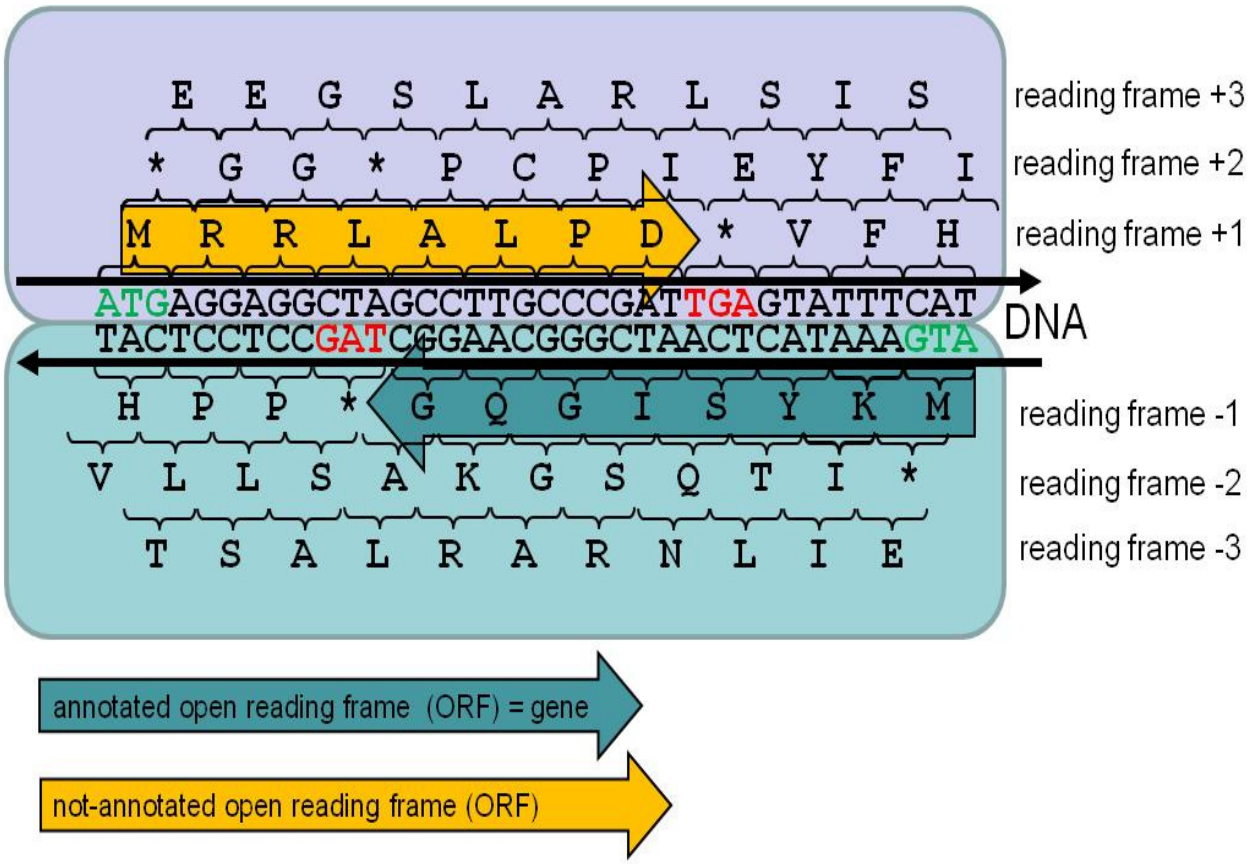

Figure 5. Illustration of the DNA double strand with its six reading frame

structural units in cells. To build a protein, genes are transcribed into (messenger) RNA, which is then translated to a sequence of amino acids - the building blocks of proteins. The genes themselves are defined by a start and stop codon (a triplet of bases) in one reading frame of the DNA. A reading frame is a sequence of codons and depending on whether a start codon begins at the first, second or third base to continue in codons (triplets), there exist three corresponding reading frames on one strand and another three on the other strand. All substrings within one reading frame which start with a start codon and end with a stop

codon are called open reading frames (abbreviated ORF). There are many ORFs and since there are six reading frames ORFs can overlap each other at one locus (see Figure 5 for an illustration of the basic terms). If a ORF is known (by experimental evidence) or assumed (by automated gene prediction) to encode a protein, it is regarded as a gene.

The challenge with RNA-seq data analysis is that typically, the size of a genome sequences is large, ranging from hundreds to billions of base pairs. For instance, the longest known vertebrate genome sequence has 130 billion base pairs. Even smaller genomes such as bacteria genomes can contain up to 13 million base pairs. The large data volume makes it impossible to perform

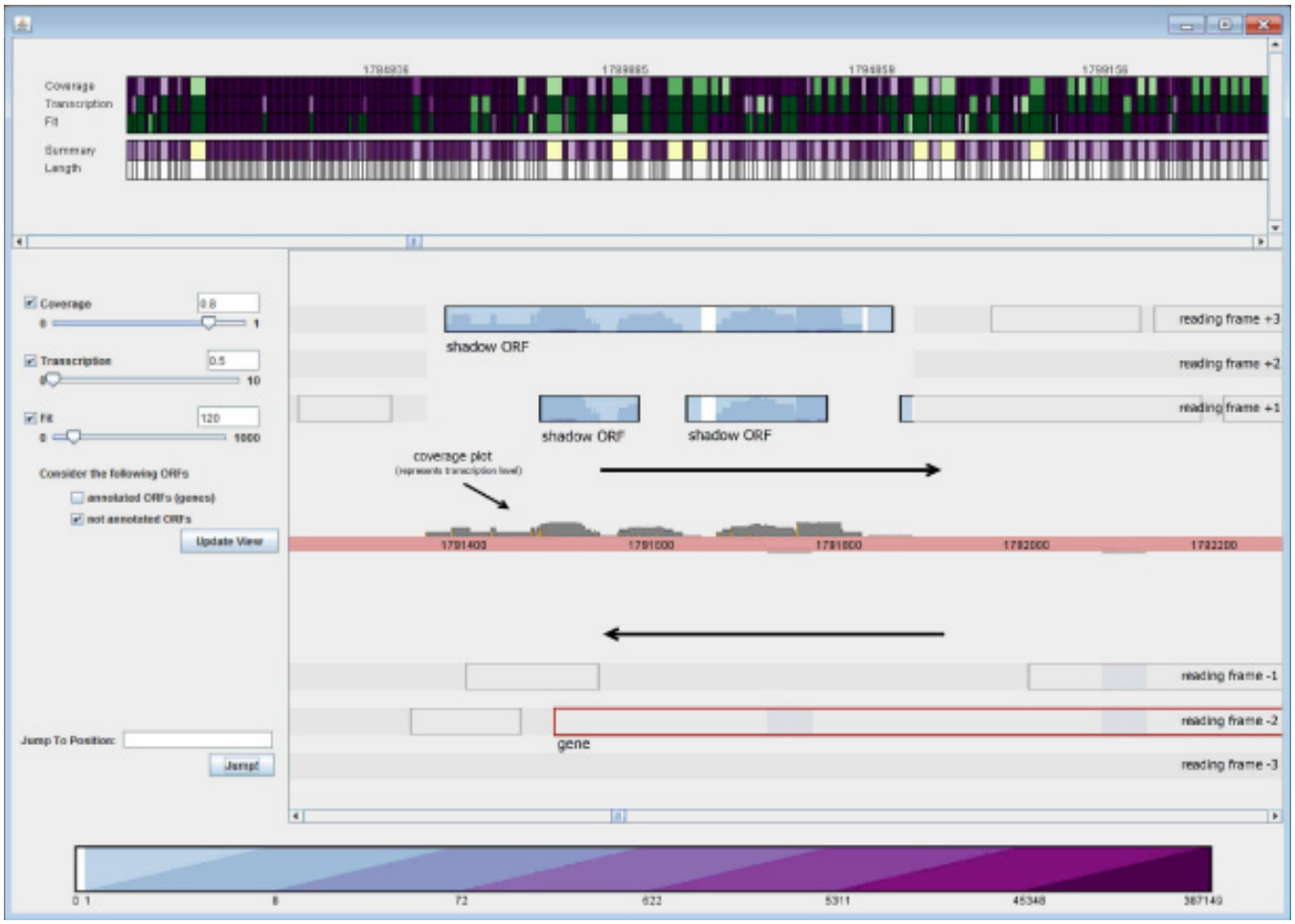

Figure 6. Screen shot of a visual analytics tool to analyze RNA-seq data from Next-Generation-Sequencing [14]. The GUI consists of three panels: the overview panel (top), the parameter setting panel (left) and the genome view panel (middle). The overview panel allows adjustment of filter parameters and navigation of the whole data set. The sliders at the parameter setting panel can be used to filter the data to reduce searching time while looking for interesting shadow ORFs (gene candidates). The genome view is split into the forward strand of the DNA (above the pink bar in the middle) and the reverse strand of the DNA (below the pink bar). For each strand three readings frames are shown, which contain genes and shadow ORFs (gene candidates which are of interest in this application). The data (coverage plot) is shown as a bar chart above and below the pink bar in the middle, this data is also encode by pseudo two-tone coloring [12] in the rectangles representing genes and shadow ORFs. The corresponding color-map is shown at the bottom of the GUI

analysis tasks by manually checking through all the ORFs. On the other hand since current technologies can only sequence small fragments, sequences are fragmented, which results among other reasons in uneven signals for a gene. Such noise and uncertainty make it impossible to apply automated analysis methods on the data without human intervention. Especially when determining the correct ORF among several overlapping ORFs, human inspection and visualizations are needed to help biologists

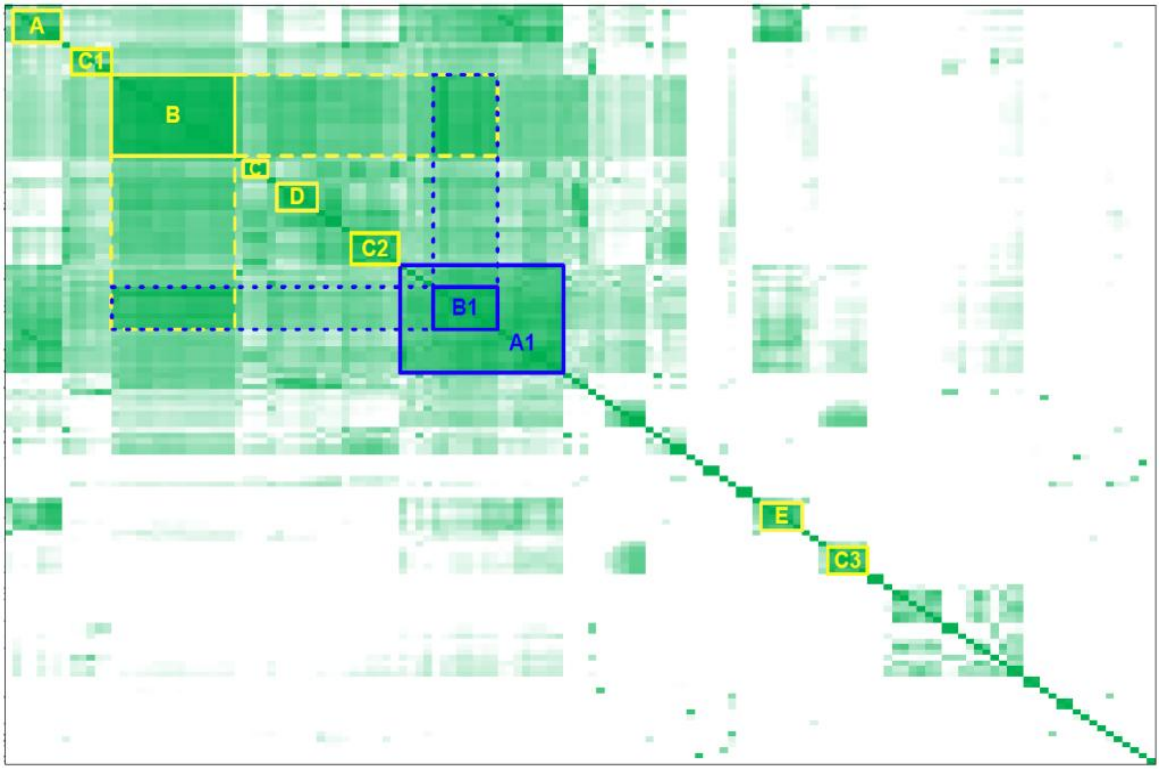

Figure 7. Correlation matrix for feature selection

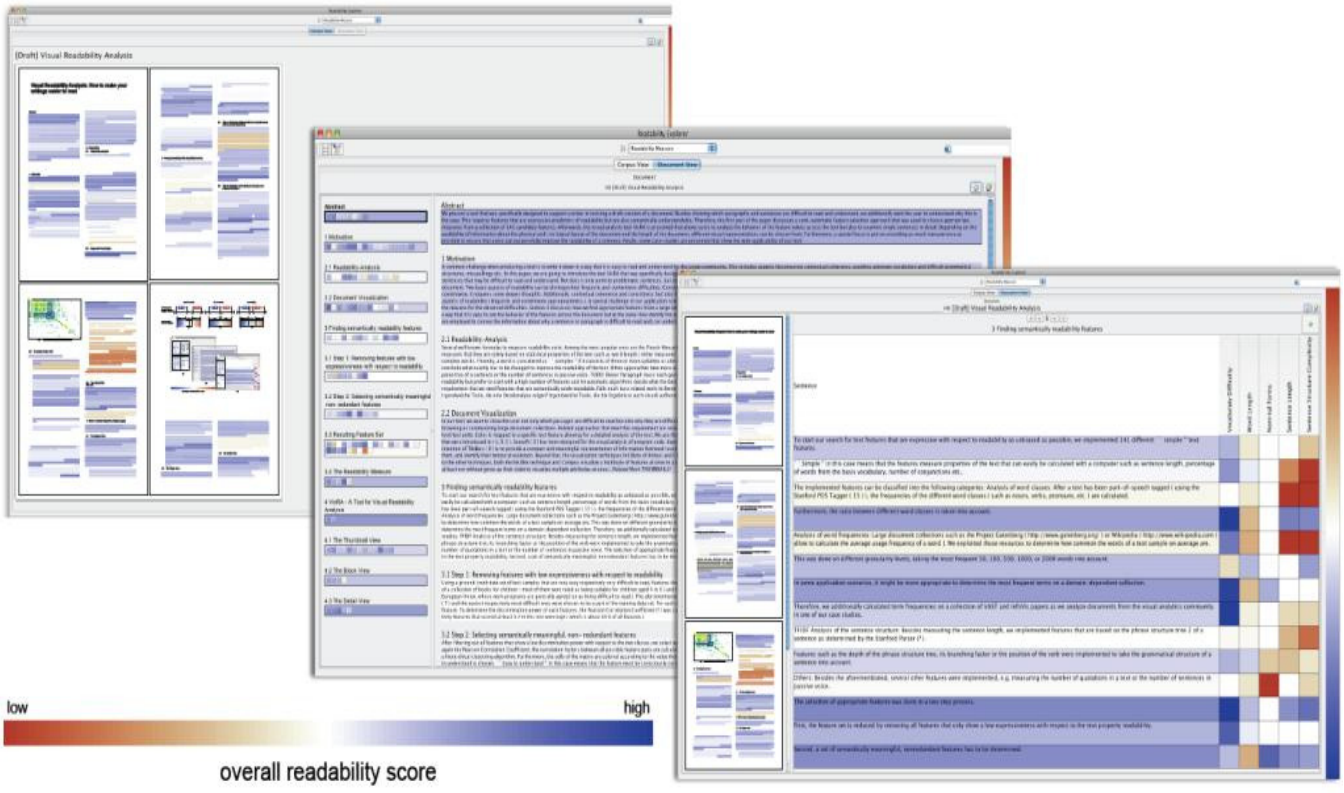

Figure 8.Screenshot of the VisRA tool on 3 different aggregation levels
(a) Corpus View (b) Block View (c) Detail View

filter out irrelevant information and narrow down their search. In [14] we developed a visual analytics tool for identifying overlapping genes in bacteria using RNA-seq data form NGS experiments. An interestingness function based on some important parameters is implemented to filter for interesting candidates. The parameters of the interestingness function can be adjusted by users dynamically at run time based on their findings. A overview representation helps to adjust the parameters. Each row represents a gene or shadow ORF and color encodes if its values are above or below the threshold. Hue represents the distance to the threshold. See figure 6 for a screen shot of the tool and an explanation of its elements.

4.4 Visual Readability Analysis

A common challenge of writing is to write in an easy-to-understand way. During the past years, a large number of features have been proposed to measure the readability of a text document ([7] and [9], for instance) and tools have been developed to help authors evaluate the readability of their documents and highlight the difficult-to-read parts [11]. However, knowing which parts of the document are difficult to read does not solve the whole problem; the author also needs to understand why this is the case.

The readability of a document is often influenced by both linguistic and content-wise difficulties. To provide the user with detailed information about why a sentence or a paragraph is difficult to read, readability measures that are both semantically rich and expressive are needed. Given the large number of features that are proposed, it is difficult to select the best features for evaluating a piece of text based on a given task. Sometimes common sense or expert knowledge is used to determine the right features. However, with such an approach it easily happens that features that do have a high expressiveness but are not commonly associated with the task are ignored. On the other hand, fully automatic feature selection techniques may end up with features that are semantically difficult to understand.

VisRA [10] is a visual analysis tool that takes a semi-automatic feature selection approach to help choose appropriate measures from a large selection of candidate readability measures and display the readability of the document in effective visual forms. First the features that do not have any discriminative power are discarded. This narrows down the search space for feature selection. A correlation Matrix (see Figure 7) is then used to help the user to remove redundant features and select the subset of features that are semantically meaningful. The tool allows the user to analyze the selected feature values across the text and within single sentences. The system puts special emphasis on providing as much transparency as possible to ensure that the user can improve the readability of a document. Figure 8 shows the three levels of details generated by the VisRA system: corpus view, block view and detail view. The corpus view can be used for navigating through the text and identifying passages that are in need for revision. The block view and the detail view show the readability of the text at different levels of details and the individual readability score based on different measures.

5. Conclusions

Nearly all grand challenge problems of the 21st century, such as climate change, the energy crisis, the financial crisis,the health and security challenges, require the analysis of very large and complex datasets, which can be done neither by the computer nor the human alone. Visual analytics is a young active science field that comes with a mission of empowering people to find solutions for complex problems from large complex datasets. By tightly integrating human intelligence and intuition with the storage and processing power of computers, many recently developed visual analytics solutions successfully help people in analyzing large complex datasets in different application domains. However, challenges remain in many disciplines of visual analytics. In this paper, we highlight some of the most important challenges and use some real world applications to show the potential of applying visual analytics techniques to help people synthesize information from heterogeneous sources and derive insight from large data sets.

References

[1] Andrienko, G., Andrienko, N., et al. (2009). Analysis of community-contributed space- and time referenced data (by example of panoramio photos). *In*: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 540–541.

[2] EMM. (2012). http://emm.newsbrief.eu/overview.html, online, accessed: 20th-August.

[3] HGP. (2012). http://www.genome.gov/, online, accessed: 20th-August.

[4] Keim, D. A., Andrienko, G., et al. (2008). Visual analytics: Definition, process, and challenges. Information Visualization, p. 154–175.

[5] Keim, D. A., Kohlhammer, J., et al., editors. (2010). Mastering the Information Age-Solving Problems with Visual Analytic. Eurographics.

[6] Kiernan, J., Terzi, E. (2009). Eventsummarizer: a tool for summarizing large event sequences. *In*: Proceedings of the 12th International Conference on Extending Database Technology, p. 1136–1139.

[7] Kincaid, J. P., Fishburn, R. P., et al. (1975). Derivation of new readability formulas for navy enlisted personnel. Research branch report 8-75, Naval Air Station Memphis.

[8] Krstajic, M., Bertini, E., et al. (2011). Cloudlines: Compact display of event episodes in multiple time-series. IEEE Transactions on Visualization and Computer Graphics, 17, 2432–2439.

[9] McLaughlin, H. G. (1969). Smog grading - a new readability formula. *Journal of Reading*, 12 (8) 639–646.

[10] Oelke, D., Spretke, D., et al. (2010). Visual readability analysis: How to make your writings easier to read. *In*: Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST '10), p. 123–130.

[11] ORT. (2012). http://www.online-utility.org, online, last accessed 20th-august.

[12] Saito, T., Miyamura, H., et al. (2005). Two-tone pseudo coloring: Compact visualization for one-dimensional data. *In*: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization, INFOVIS '05, p. 23, Washington, DC, USA. IEEE Computer Society.

[13] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *In*: Proceedings of the IEEE Symposium on Visual Languages, p. 336–343, Washington, IEEE Computer Society Press.

[14] Simon, S., Oelke, D., et al. (2011). Visual analysis of next-generation sequencing data to detect overlapping genes in bacterial genomes. In Proceedings of 1st IEEE Symposium on Biological Data Visualization, p. 47–54

[15] Thomas, J., Cook, K. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Ctr.