Visual Analytics: Combining Automated Discovery with Interactive Visualizations

Daniel A. Keim, Florian Mansmann, Daniela Oelke, and Hartmut Ziegler

University of Konstanz, Germany
first.lastname@uni-konstanz.de,
WWW home page: http://infovis.uni-konstanz.de

Abstract. In numerous application areas fast growing data sets develop with ever higher complexity and dynamics. A central challenge is to filter the substantial information and to communicate it to humans in an appropriate way. Approaches, which work either on a purely analytical or on a purely visual level, do not sufficiently help due to the dynamics and complexity of the underlying processes or due to a situation with intelligent opponents. Only a combination of data analysis and visualization techniques make an effective access to the otherwise unmanageably complex data sets possible.
Visual analysis techniques extend the perceptual and cognitive abilities of humans with automatic data analysis techniques, and help to gain insights for optimizing and steering complicated processes. In the paper, we introduce the basic idea of Visual Analytics, explain how automated discovery and visual analysis methods can be combined, discuss the main challenges of Visual Analytics, and show that combining automatic and visual analysis is the only chance to capture the complex, changing characteristics of the data. To further explain the Visual Analytics process, we provide examples from the area of document analysis.

1 Introduction

The information overload is a well-known phenomenon of the information age, since our ability to collect and store data is increasing at a faster rate than our ability to analyze it. In numerous application areas fast growing data sets develop with ever higher complexity and dynamics. The analysis of these massive volumes of data is crucial in many application domains. For decision makers it is an essential task to rapidly extract relevant information from the immense volumes of data. Software tools help analysts to organize their information, generate overviews and explore the information in order to extract potentially useful information. Most of these data analysis systems still rely on visualization and interaction metaphors which have been developed more than a decade ago and it is questionable whether they are able to meet the demands of the ever-increasing masses of information. In fact, huge investments in time and money are often lost, because we lack the possibilities to make proper use of the available data. The basic idea of Visual Analytics is to visually represent the information, allowing

the human to directly interact with the data to gain insight, draw conclusions, and ultimately make better decisions. The visual representation of the information reduces complex cognitive work needed to perform certain tasks. "People use visual analytics tools and techniques to synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data ... to provide timely, defensible, and understandable assessments" [1].

The goal of Visual Analytics research is to turn the information overload into an opportunity. Decision-makers should be enabled to examine this massive information stream to take effective actions in real-time situations. For informed decisions, it is indispensable to include humans in the data analysis process and combine their flexibility, creativity, and background knowledge with the enormous storage capacity and the computational power of today's computers. The specific advantage of Visual Analytics is that decision makers may focus their full cognitive and perceptual attention on the decision, while allowing them to apply advanced computational methods to make the discovery process more effective.

The rest of this paper is structured as follows: Section 2 defines Visual Analytics, discusses related research areas, and presents a model of the Visual Analytics Process. In Section 3, we discuss the major technical challenges of the field. To foster a deeper understanding of Visual Analytics, Section 4 details examples of how visual and automatic methods can be used for an advanced interactive document analysis. Finally, Section 5 summarizes the key aspects of our paper.

2 Visual Analytics

In this section we will discuss Visual Analytics by defining it, by listing related research areas, and by presenting a model of the Visual Analytics Process.

2.1 Definition

According to [1], Visual Analytics is the science of analytical reasoning supported by interactive visual interfaces. Today, data is produced at an incredible rate and the ability to collect and store the data is increasing at a faster rate than the ability to analyze it. Over the last decades, a large number of automatic data analysis methods have been developed. However, the complex nature of many problems makes it indispensable to include human intelligence at an early stage in the data analysis process. Visual Analytics methods allow decision makers to combine their human flexibility, creativity, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems. Using advanced visual interfaces, humans may directly interact with the data analysis capabilities of today's computer, allowing them to make well-informed decisions in complex situations.

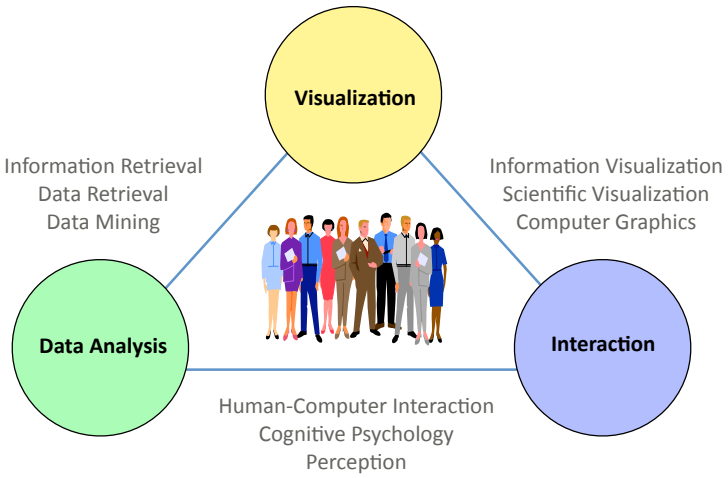


Fig. 1. Research Areas Related to Visual Analytics

2.2 Related Research Areas

Visual Analytics can be seen as an integral approach combining visualization, human factors, and data analysis. Figure 1 illustrates the research areas related to Visual Analytics. Besides visualization and data analysis, especially human factors, including the areas of cognition and perception, play an important role in the communication between the human and the computer, as well as in the decision-making process. With respect to visualization, Visual Analytics relates to the areas of Information Visualization and Computer Graphics, and with respect to data analysis, it profits from methodologies developed in the fields of information retrieval, data management & knowledge representation as well as data mining.

2.3 The Visual Analytics Process

The Visual Analytics Process combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. Figure 2 shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the Visual Analytics Process.

In many application scenarios, heterogeneous data sources need to be integrated before visual or automatic analysis methods can be applied. Therefore, the first step is often to preprocess and transform the data to derive different representations for further exploration (as indicated by the *Transformation* arrow in Figure 2). Other typical preprocessing tasks include data cleaning, normalization, grouping, or integration of heterogeneous data sources.

After the transformation, the analyst may choose between applying visual or automatic analysis methods. If an automated analysis is used first, data mining

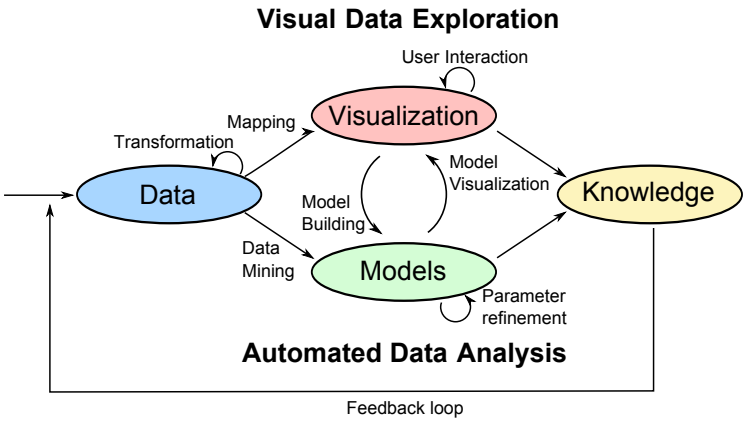


Fig. 2. The *Visual Analytics Process* is characterized through interaction between data, visualizations, models about the data, and the users in order to discover knowledge.

methods are applied to generate models of the original data. Once a model is created the analyst has to evaluate and refine the models, which can best be done by interacting with the data. Visualizations allow the analysts to interact with the automatic methods by modifying parameters or selecting other analysis algorithms. Model visualization can then be used to evaluate the findings of the generated models. Alternating between visual and automatic methods is characteristic for the Visual Analytics process and leads to a continuous refinement and verification of preliminary results. Misleading results in an intermediate step can thus be discovered at an early stage, leading to better results and a higher confidence. If a visual data exploration is performed first, the user has to confirm the generated hypotheses by an automated analysis. User interaction with the visualization is needed to reveal insightful information, for instance by zooming in on different data areas or by considering different visual views on the data. Findings in the visualizations can be used to steer model building in the automatic analysis. In summary, in the Visual Analytics Process knowledge can be gained from visualization, automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts.

The Visual Analytics Process aims at tightly coupling automated analysis methods and interactive visual representations. The classic way of visually exploring data as defined by the Information Seeking Mantra ("Overview first, Zoom/Filter, Details on demand") [2] therefore needs to be extended to the Visual Analytics Mantra [3]:

"Analyze First -
Show the Important -
Zoom, Filter, and Analyze Further -
Details on Demand"

With massive data sets at hand all three steps of the Information Seeking Mantra are difficult to implement. An overview visualization without losing interesting patterns is difficult to create, since the amount of pixels of the display does not keep pace with the increasing flood of data. In Visual Analytics, it is therefore not sufficient to just retrieve and display the data using a visual metaphor; it is rather necessary to analyze the data according to its value of interest, show the most relevant aspects of the data, and at the same time provide interaction models, which allow the user to get details of the data on demand.

3 Challenges of Visual Discovery

With information technology becoming a standard in most areas in the past years, more and more digital information is generated and collected. As the amount of data is continuously growing and the amount of pixels on the display remains rather constant, the huge amount of data to be visualized exceeds the limited amount of pixels of a display by several orders of magnitude. One key challenge of Visual Analytics is therefore *scalability* as it determines the ability to process large datasets in terms of computational overhead. In particular, since we are dealing with visualization techniques, the visual scalability of the techniques has to be considered, which is defined as the capability of visualization tools to effectively display large data sets in terms of either the number or the dimension of individual data elements [4]. While relying on increased hardware performance to cope with larger and larger problems, researchers need to design more effective Visual Analytics algorithms to bring this data onto the screen in an appropriate way.

Tremendous streams of time related or real time data are generated by dynamic processes, arising in business, networks, or telecommunications. Examples are sensor logs, web statistics, network traffic logs, or atmospheric and meteorological records. Analyzing these *data streams* is an important challenge, since the sheer amount of data does often not allow to record all data at full detail. This results in the need for effective compression and feature extraction to manage and access the data. Furthermore, real-time requirements put an additional burden upon the application developers. To enable quick identification of important information and timely reaction to critical process states or alarming incidents, analysis techniques and metaphors need to be developed, which render the user capable of analyzing real time data streams by presenting the results instantly in a meaningful and intuitive way.

To be capable of accessing information from a number of different sources, real-world applications require scalable methods for the *synthesis of heterogeneous types of data*. The heterogeneous data sources may include collections of vector data, strings, text documents, graphs, or multimedia objects. Integrating these data sources touches a number of fundamental problems in decision theory, information theory, statistics, and machine learning, evidently posing a challenge for Visual Analytics, too. The focus on scalable and robust methods for fusing complex heterogeneous data sources is thus key to a more effective

analysis process. Computational biology is one such application domain where the human genome, for example, is accompanied by real-valued gene expression data, functional annotation of genes, genotyping information, a graph of interacting proteins, equations describing the dynamics of a system, localization of proteins in a cell, and natural language documents in the form of papers describing experiments or partial models.

Visual Analytics can also help to close the *Semantic Gap*. Since humans are the ultimate instance for defining semantics, Visual Analytics may significantly improve the way semantic definitions are obtained and refined. In particular, methods from semantics research may capture associations and complex relationships within the data sets to support decision-centered visualization. While ontology-driven techniques and systems have already started to enable new semantic applications in a wide span of areas, further research is necessary to increase our capabilities for creating and maintaining large domain ontologies and automatic extraction of semantic meta data, since the integration of different ontologies to link various datasets is hardly automated yet. Research challenges arise from the size of ontologies, content heterogeneity, and link analysis over ontology instances or meta data. New Visual Analytics methods to resolve semantic heterogeneity and discover complex relationships are thus needed.

Finally, *evaluation* as a systematic determination of merit, worth, and significance of a technique or system is essential to the success of Visual Analytics. Different aspects need to be considered when evaluating a system, such as functional testing, performance benchmarks, measurement of the effectiveness of the display, economic success, user studies, assessment of its impact on decision-making, etc. Note that not all of these aspects are orthogonal nor can they always be applied. Since Visual Analytics deals with unprecedented data sizes, many developed applications contain novel features to support a previously unsolvable analysis task. In such a case, the lack of a competing system turns a meaningful evaluation into a challenge in itself.

4 Example Application: Visual Document Analysis

Document Analysis is an area in which the need for visual analysis techniques is quite obvious. Large amounts of information are only available in textual form (e.g. books, newspapers, patents, service reports, etc.). But often these valuable resources are not used, because reading and analyzing the documents would take too much effort. Take for example a company's need to know the public opinion about one of its products and especially about rumors regarding that product. Knowing about such rumors is important to be able to quickly react to undesired developments and to effectively influence the public opinion in a favorable way. The Internet is a great place for understanding the public opinion since nowadays a significant percentage of the population participates in writing blogs, commenting on products at merchant sites, stating their opinions in forums, etc. And people read other people's comments to get information and form their opinion. With current search engines, however, it is not easy to

find the relevant information related to the public opinion about a company's product, since search engines usually return millions of hits with only a small percentage being relevant to the task.

The example shows that it is impossible for a human to find and analyze all the relevant documents. On the other hand, an automatic semantic analysis of the documents is still infeasible today due to a) the impressive flexibility and complexity of natural language as well as b) the need to semantically interpret the content. The challenge that researchers try to tackle with Visual Analysis techniques is how to allow the human and computer to effectively work together to bridge the Semantic Gap.

Text can be analyzed on different abstraction levels:

- statistical level (e.g. frequencies of (specific) words, average sentence length, number of tokens or types, etc.)
- structural level (structural components of a document, such as header, footer, title, abstract, etc.)
- syntactical level (principles and rules for constructing sentences)
- semantic level (linguistic meaning)
- pragmatic level (meaning in context; consequence for actions)

The higher the abstraction level the more difficult it is for the computer to appropriately analyze a text. Counting words and characters as done at the statistical level is a simple task which can easily be performed by the computer. The identification of the structure of a document and the analysis of the syntax is already more challenging but can still be computationally approached (see e.g. the techniques presented in [5] [6]). Analyses on the semantic and pragmatic level are much more challenging. The idea of Visual Analytics is to let the human and the computer cooperate in solving the task. Humans contribute background knowledge, interpretation, and semantic analysis of the text whereas the computer supports the human analysts in the best possible way to enable them to deal with large data sets, e.g. by performing the time-consuming preprocessing and filtering steps.

4.1 Quasi-semantic Document Properties

The vision for automatic document analysis is to teach the computer to understand a document in a way similar to humans including its semantic and pragmatic aspects. Since this goal seems to be too ambitious at the current state of research, we start by teaching the computer to analyze a document with respect to one semantic aspect at a time. This task is relevant in many real application scenarios. Often large amounts of documents have to be analyzed with respect to a certain analysis question. Examples for such document analysis questions include:

- What is the public opinion regarding a product / a politician / a "hot" news topic, etc. that is expressed in news articles, blogs, discussion groups, etc. on the Internet?

- How trustworthy are the statements?
- How much emotion content is contained in the documents (e.g. hate in terrorist webpages)?

We call the document property that is central to the considered document analysis question a *quasi-semantic property*. We define quasi-semantic properties as higher-level document properties that capture one semantic aspect of a document (e.g. positive / negative statements with respect to a given product name). Most quasi-semantic properties cannot be measured directly. Nevertheless, combinations of low-level features (i.e. statistical, structural and syntactical features) can be used to approximate quasi-semantic properties of the documents, which help to bridge the semantic gap. The challenge is how to determine the best combination of low-level features to approximate such higher-level document properties.

4.2 Opinion Analysis

Figure 3 shows a set of documents that have been analyzed with respect to a quasi-semantic property that tries to assess the positive or negative opinion expressed in the documents. To automatically assess the polarity of a sentence we counted the number of opinion signal words. The signal words were given in form of two lists that contain adjectives, verbs and nouns (such as "bad", "problem", "wonderful", etc.) that hint at a subjective statement and its polarity. The algorithm can easily be improved by taking context dependent signal words or negation into account (cf. [7] and [8]). For illustrative purposes, we use the title pages of the November 2007 issues of *The Telegraph* as text corpus. The figure shows that there are some articles that are completely negative (e.g. the article in the lower right corner) and others that are mostly positive (such as the articles about the Queen in the 1st column, 3rd row). Interestingly, there are also articles with quite mixed opinions or with a sudden change in polarity (for example, the first article in the last column, 1st row or the lower left article in the 4th column, 1st row). The example demonstrates that by combining automatic and visual methods it becomes possible to quickly analyze a document corpus with respect to a quasi-semantic property without reading it.

4.3 Authorship Attribution

Our second case study shows how Visual Analytics techniques can be used to analyze the discrimination power of low-level features that are commonly used in authorship attribution. Given some documents with known authorship the task of authorship attribution is to assign a document with unknown authorship to the author that has written it. Thus, in this case the quasi-semantic property that we would like to measure is the writing style of an author.

In previous work we focused on the development of techniques that support the analysis of low-level features and thus can be used to find (combinations of)

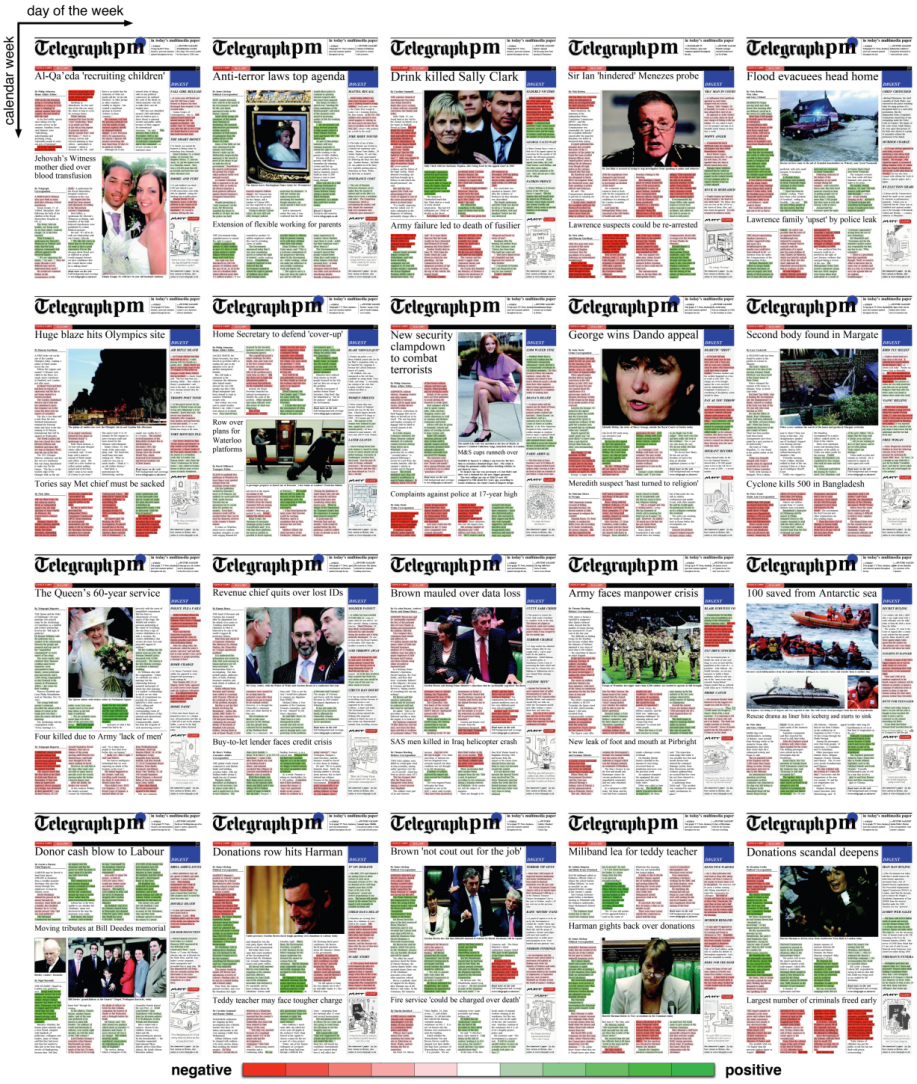


Fig. 3. Title pages of *The Telegraph* in November 2007. The text has been analyzed with respect to a quasi-semantic property 'that tries to assess the positive or negative opinion expressed in the documents. Sentences with positive statements are highlighted in green, the ones with negative statements in red, respectively. The degree of positiveness or negativeness is denoted by the intensity of the color. (courtesy of *The Telegraph*)

low-level features that are able to approximate a desired quasi-semantic property. In fully automatic document analysis often just a single feature value is calculated per document. With the use of visualization techniques, it is possible to extract a sequence of feature values and present it to the user as a characteristic fingerprint for each document. By doing this it is possible to analyze the development of the values across the document in detail. Figure 4 shows our Literature Fingerprinting technique which was first presented in [9]. In Figure 4, the technique is applied to several books of Jack London (first row of each subfigure) and Mark Twain (last three rows of each subfigure). Each pixel represents a text block of 10,000 words and the pixels are arranged from left to right and top to bottom as they appear in the sequence of the book. Neighboring blocks have an overlap of 9,000 words to obtain a continuous and split-point independent representation. Color is mapped to the feature value, ranging from blue for high feature values to red for low feature values. In this example the values of five different features have been calculated:

- the average sentence length
- the frequencies of specific function words; the resulting high-dimensional feature vectors are projected into low-dimensional space using a Principal Component Analysis and the first and second dimension are visualized in Figures 4(c) and 4(d).
- three vocabulary measures, namely Hapax Legomena Index, Hapax Dislegomena Index and Simpson's Index which are calculated as follows:
  Hapax Legomena Index (R):

$$R = \frac{100 \log N}{1 - V_1/V}$$

  Hapax Dislegomena Index (D):

$$D = \frac{V_2}{V}$$

  Simpson's Index (S):

$$S = \frac{\sum_{r=1}^{\infty} r(r-1)V_r}{N(N-1)}$$

  where N = the number of tokens V = the number of types V_r = the number of lexical units that occur exactly r times

Please refer to [10] for an overview of the different features that are used for authorship attribution.

Each subfigure shows visualizations of all documents for one specific low-level feature. If the feature is able to discriminate between the two authors, the books in the first row (books by Jack London) have to be different from the ones in the last three rows (books by Mark Twain). It can easily be seen that there are some low-level features for which this is largely true, e.g. average sentence length in Figure 4(a) but also Simpson's Index in Figure 4(b). Others do not seem to

have any discrimination power with respect to the two authors at all (e.g. Hapax Dislegomena which is depicted in Figure 4(f)). Interestingly, there is one book of Mark Twain that sticks out in many visualization, namely *The Adventures of Huckleberry Finn* (middle book in the middle row of the books by Mark Twain). The writing style of this book seems to be totally different from all the other books of Mark Twain.

This case study shows a small example of how Visual Analytics may help in better solving complex analysis tasks. The visual analysis enables the analyst to detect problems with the low-level feature used and adapt the similarity measures to make the authorship attribution more effective. While the example clearly shows the advantage of Visual Analytics it is only a first step toward a Visual Document Analysis system which tightly integrates automated document analysis and interactive document exploration capabilities.

5 Conclusions

Since data volumes are increasing at a faster pace than our ability to analyze them, there is a pressing need for automatic analysis methods. However, most automatic analysis methods require a well-defined problem and often return large and complex models. Visual Analytics turns the information overload problem into an opportunity by integrating interactive data exploration with advanced knowledge discovery algorithms.

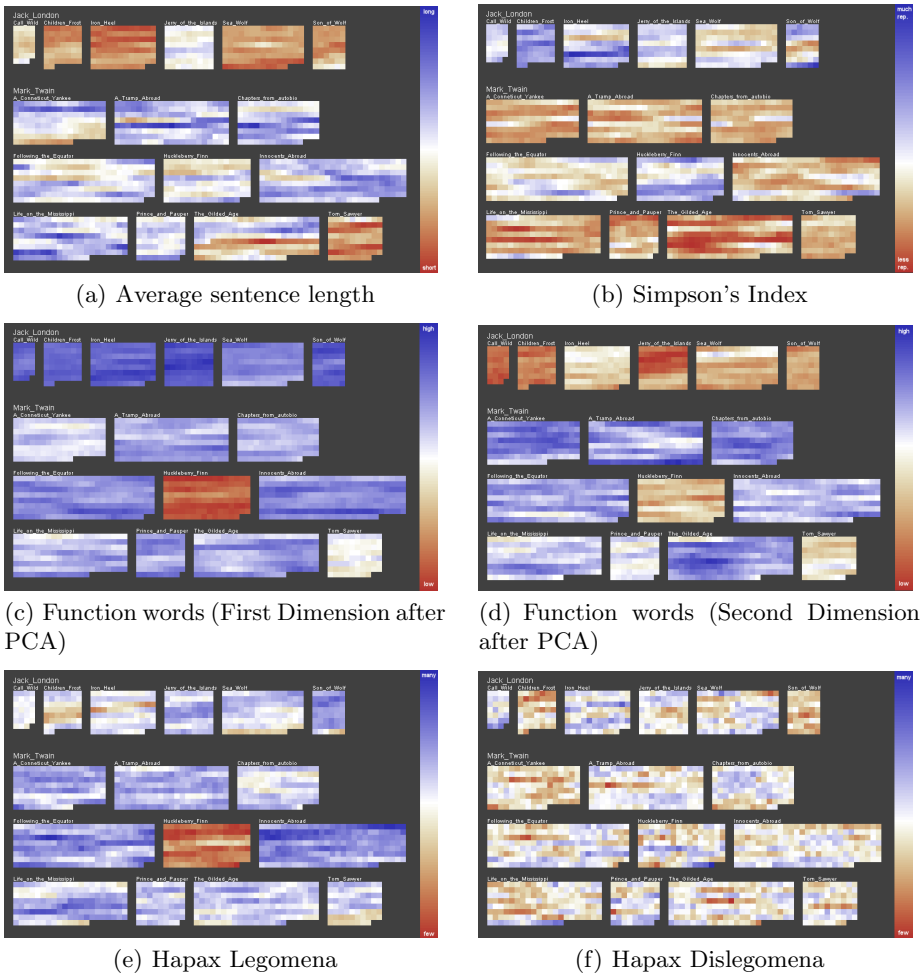
In this paper, we motivate and define Visual Analytics, present a model of the Visual Analytics Process for a deeper understanding of how methods from visual data exploration and information mining can be combined to gain insights into large and complex datasets. The paper sketches the main challenges of Visual Analytics and describes why these challenges are difficult to solve. In particular, we give a demonstrative example of how Visual Analytics methods can help to gain insights in document analysis with an application to the authorship attribution problem.

Acknowledgement

We thank Jörn Schneidewind for helpful comments on the manuscript.

References

1. Thomas, J., Cook, K.: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press (2005)
2. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: IEEE Symposium on Visual Languages. (1996) 336–343
3. Keim, D.A., Mansmann, F., Schneidewind, J., Ziegler, H.: Challenges in visual data analysis. In: Information Visualization (IV 2006) ,Invited Paper, July 5-7, London, United Kingdom, IEEE, IEEE Press (2006)



(a) Average sentence length

(b) Simpson's Index

(c) Function words (First Dimension after PCA)

(d) Function words (Second Dimension after PCA)

(e) Hapax Legomena

(f) Hapax Dislegomena

Fig. 4. *Literature Fingerprinting Technique (see [9]).* Instead of calculating a single feature value per document, a sequence of feature values is extracted and presented to the user as a characteristic fingerprint for each document. In the example above, the technique is used to analyze the discrimination power of text features for authorship attribution. Each pixel represents the feature value for one text block and the grouped pixels belong to one book. The different feature values are mapped to color. If a feature is able to discriminate between the two authors, the books in the first row (that have been written by J. London) are visually different from the remaining books (written by M. Twain). Each subfigure shows the visualization of the values of one specific low-level feature that is commonly used for authorship attribution. It can easily be seen that not all features are able to discriminate between the two authors. Furthermore, it is interesting to observe that the book *Huckleberry Finn* (middle book in the middle column of the books of M. Twain) sticks out in a number of features as if it was not written by Mark Twain.

4. Eick, S.G.: Visual scalability. Journal of Computational & Graphical Statistics (March 2002)
5. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2003) 423–430 http://nlp.stanford.edu/software/lex-parser.shtml.
6. Hadjar, K., Rigamonti, M., Lalanne, D., Ingold, R.: Xed: a new tool for extracting hidden structures from electronic documents. In: International Workshop on Document Image Analysis for Libraries. (2004) 212–224
7. Oelke, D., Bak, P., Keim, D., Last, M., Danon, G.: Visual evaluation of text features for document summarization and analysis. In: IEEE Symposium on Visual Analytics and Technology (VAST 2008). (2008) to appear.
8. Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: WSDM '08: Proceedings of the international conference on Web search and web data mining, New York, NY, USA, ACM (2008) 231–240
9. Keim, D., Oelke, D.: Literature fingerprinting: A new method for visual literary analysis. In: IEEE Symposium on Visual Analytics and Technology (VAST 2007). (2007)
10. Holmes, D.I.: Authorship Attribution. Computers and the Humanities **28** (1994) 87–106