Exploratory Visual Text Analytics in the Scientific Literature Domain

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der
Universität Stuttgart
zur Erlangung der Würde eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Abhandlung

Vorgelegt von

Florian Manuel Heimerl

aus Kronach

Hauptberichter:	Prof. Dr. Thomas Ertl
Mitberichter:	Prof. Christopher Collins, Ph.D.
Tag der mündlichen Prüfung:	08. Dezember 2016

Institut für Visualisierung und Interaktive Systeme
der Universität Stuttgart

2017

Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Acknowledgment

While working on the projects for this thesis, I was lucky to have had the opportunity to work and share ideas and discussions with many wonderful people. First and foremost, I am deeply grateful to my supervisor, Thomas Ertl, for giving me the opportunity to be part of his great group, for his guidance, and his unprecedented support. For his interest in and feedback on my work I would like to thank Christopher Collins. It was a pleasure having him as my external examiner. I am also grateful to Steffen Koch for introducing me to the craft of visualization, and for his guidance and many fruitful discussions. During the past five years I was able to work with a wonderful team of incredible talented people that I am very proud to have been part of. I am grateful to Kuno Kurzhals, my office partner for five years, and my colleagues Qi Han, Markus John, Robert Krüger, Dennis Thom, Harald Bosch, Marcel Hlawatsch, Michael Burch, Steffen Lohmann, Michael Wörner, Dominik Herr, Tanja Blascheck, and many other members of VIS and VISUS for developing ideas together, working on joint projects, and the fun and productive time that we had. I would also like to thank my project collaborators Hinrich Schütze and Charles Jochim for the fruitful collaboration and the *Deutsche Forschunsgemeinschaft* (DFG) for funding my research. Some of the projects I was involved in during my work on this thesis were done jointly with external collaborators who I would like to thank, most notably Paolo Federico, Silvia Miksch, and Petra Isenberg. I also want to thank all students that I had the pleasure of supervising or working with. These include Moataz Abdelfattah, Edwin Püttmann, Eduard Marbach, Constanze Sudra, Cyrill Fabian Bopp, and Simon Lange. Many thanks to Martin Falk who created the beautiful layout this thesis is typeset in. I am particularly grateful to my family and partner for their love and support, and their patience, especially during the final phase of writing the thesis.

Abstract

Digitization has become the prevailing mode of producing and storing data. Driven by technological progress, this trend yields a deluge of digitally stored information. Apart from machine understandable, often numerical, data with a clearly defined semantics, this includes text data, on which we focus, but also audio or video recordings. The ubiquity of hand-held devices that allow access to this data anywhere at any time, such as smart phones, tablet computers, and e-book readers, has further spurred this trend. Magazines, newspapers, and other traditional print products are nowadays mostly created, edited, and delivered in digital form. In addition, the entire production process of scientific journals and conference proceedings is being handled digitally. Data generated and stored as a result of this offer the grand opportunity of significantly speeding up access to information and knowledge. For scientific literature, this means support for long existing problems, such as related work search, trend analysis in new disciplines, and evaluating important authors. While scientific literature is similar to other documents in many respects, it differs as for its technical language, and wealth of meta data.

Although automatic processing of text data is possible to a certain extent, autonomously extracting and organizing all information from natural language texts is unfeasible. Interpreting texts needs human intuition and world knowledge. Visual analytics is a young research field that combines the ability of computers to recognize patterns in large amounts of data and the knowledge and linguistic capabilities of human users. It integrates interactive visualization to interact with abstract data representations, and thus provides a basis for sensemaking and human reasoning. In addition, it comprises data mining and aggregation methods that help to process and abstract data. Through interaction, users are enabled to steer and adapt them to their analysis needs. Visual analytics approaches thus enable users to efficiently organize and handle huge amounts of text, view and analyze their contents, and make sense of them in a scalable way, without having to read every single document. This helps to derive new knowledge through the amalgamation and correlation of information hidden within such text collections.

Previous approaches for visual text analytics are often designed to support search and retrieval scenarios. These assume that users go about the analysis with previous knowledge about what type of information they are looking for, and they facilitate fast access to it. Open exploration, in contrast, that starts with vague or no objectives at all, and with little knowledge about the data sets at hand, is widely unsupported. It requires detailed, user-driven analysis loops to allow users to get an understanding of its contents and gradually develop analysis interests. The methods developed within this thesis aim to support different aspects of such

open exploration of scientific text collections. This thesis includes two novel, exploratory text analytics approaches for single text bodies as well as for large and heterogeneous sets of documents. Both extend the state-of-the-art of visual text exploration by facilitating flexible adaption of the analysis granularity. We further present two approaches that facilitate visual text classifier creation. One of them is designed for non-expert users to store the results of their exploration, and the second one supports natural language processing experts with creating adequate feature representations for accurate classification. The last approach we discuss supports the exploration of a scientific literature set by visually linking textual content and meta data. It includes the first visual method to correlate research topics with their corresponding citations and allows for their joint exploration. Although we focus specifically on the analysis of scientific documents, most of the presented methods can be applied to a large variety of documents and text types. We finally develop a taxonomy for visual text analysis approaches that identifies the joint and differing aspects of the multiple strategies that are explored within this thesis and derive open challenges for future research from it.

Zusammenfassung

Digitalisierung ist mittlerweile das vorherrschende Paradigma beim Produzieren und Speichern von Daten. Angetrieben durch den technologischen Fortschritt sorgt dieser Trend für eine unüberschaubare Menge an digital gespeicherten Informationen. Außer machinenverständlichen, oft numerischen Daten mit klar definierter Semantik, umfasst dies Textdaten, mit welchen sich diese Arbeit beschäftigt, aber auch Audio- und Videoaufnahmen. Die Allgegenwärtigkeit von mobilen Geräten wie Smartphones und Tablets, die Zugriff auf diese Daten jederzeit und an jedem Ort erlauben, heizen diesen Trend weiter an. Zeitschriften, Zeitungen und andere traditionelle Printprodukte werden heutzutage größtenteils digital erstellt, editiert und ausgeliefert. Darüber hinaus wird der komplette Produktionsprozess wissenschaftlicher Zeitschriften und Konferenzbände digital abgewickelt. Daten, die in diesem Prozess erfasst und gespeichert werden, bergen die einzigartige Gelegenheit den Zugriff zu Informationen und Wissen signifikant zu beschleunigen. Für wissenschaftliche Literatur bedeutet das Unterstützung für bereits lange existierende Herausforderungen wie die Suche nach verwandter Literatur, Trendanalyse in neuen Disziplinen und die Evaluation wichtiger Autoren. Während wissenschaftliche Literatur anderen Dokumententypen in vielen Aspekten sehr ähnlich ist, unterscheidet sie sich erheblich in Bezug auf ihre technische Sprache und die Fülle ihrer Metadaten.

Wenngleich automatische Verarbeitung von Textdaten in gewissen Grenzen machbar ist, ist die autonome Extraktion und Organisation aller in natürlichsprachlichen Texten enthaltener Information unmöglich. Die Interpretation von Texten benötigt menschliche Intuition und Weltwissen. Die Visuelle Analytik ist ein junges Forschungsfeld, das die Fähigkeit von Computern, Muster in großen Datenmengen zu erkennen mit dem Wissen und dem linguistischen Talent menschlicher Benutzer verbindet. Hierzu werden interaktive Visualisierungen verwendet, um mit abstrakten Datenrepräsentation zu interagieren. Dies schafft eine Basis für Menschen, Daten in einen sinnvollen Zusammenhang zu stellen und daraus Folgerungen abzuleiten. Zusätzlich integriert die Visuelle Analytik meist Data-Mining- und Aggregationsmethoden die dabei helfen, Daten zu verarbeiten und zu aggregieren. Mithilfe von Interaktion werden Benutzer dabei in die Lage versetzt, diese Methoden an ihre Analyseziele anzupassen und zu beeinflussen. Methoden der Visuellen Analytik ermöglichen es dabei Nutzern effizient große Mengen an Text zu ordnen und zu sichten. Dies hilft letztendlich dabei neues Wissen durch die Zusammenführung und Verknüpfung von in Textsammlungen enthaltener Information zu generieren.

Frühere Ansätze der visuellen Textanalyse wurden oft für Such- und Abfrageszenarios entwickelt. Diese gehen davon aus, das Benutzer eine Analyse mit bereits

vorhandenem Wissen über die gesuchte Information beginnen und unterstützen den schnellen Zugriff darauf. Im Gegenteil dazu wird offene Exploration, welche mit schwach definierten oder gar keinen Zielen und mit wenig Nutzerwissen über den verwendeten Datensatz startet, selten unterstützt. Eine solche Exploration benötigt kleinteilige, vom Benutzer getriebene Analysezyklen, welche es Benutzern ermöglichen ein Verständnis des Inhalts sowie allmählich ein Analyseinteresse zu entwickeln. Die Methoden, welche in dieser Arbeit entwickelt wurden, zielen auf verschiedene Aspekte einer solch offenen Exploration für Sammlungen wissenschaftlicher Literatur ab. Diese Arbeit beinhaltet zwei neuartige, explorative Textanalyseansätze für einzelne Textkörper sowie große und heterogene Dokumentenmengen. Beide erweitern den aktuellen Stand der Wissenschaft in der visuellen Textexploration durch die Unterstützung einer flexiblen Anpassung der Analysegranularität. Darüber hinaus werden zwei Ansätze präsentiert, die eine visuelle Erstellung von Textklassifikatoren ermöglichen. Einer von ihnen wurde für Nichtexperten entworfen, um die Ergebnisse ihrer Exploration festzuhalten. Der zweite unterstützt Experten der maschinellen Sprachverarbeitung dabei, geeignete Merkmalsrepräsentationen zu erstellen um eine präzise Klassifikation zu erlauben. Der letzte vorgestellte Ansatz unterstützt die Exploration einer Sammlung wissenschaftlicher Literatur indem textueller Inhalt und Metadaten visuell verknüpft werden. Er enthält die erste visuelle Methode um Forschungsthemen mit den dazugehörigen Zitationen zu verknüpfen und ihre gemeinsame Exploration zu unterstützen. Wenngleich der Fokus dieser Arbeit speziell auf der Analyse wissenschaftlicher Dokumente liegt, können die meisten der enthaltenen Methoden auf eine große Vielfalt von Dokumenten und Texttypen angewendet werden. Die Arbeit enthält weiter eine Taxonomie für visuelle Textanalyseansätze, die gemeinsame und unterschiedliche Aspekte der verschiedenen, in dieser Arbeit verfolgten Strategien, identifiziert. Hieraus werden schließlich offene Herausforderungen für zukünftige Forschung abgeleitet.

Chapter 1

Introduction

We, as humans, are immersed in language. It is our most versatile means of communication, and as such plays an outstanding role in our societies. Not only is it used as an everyday means to pass on information, stories, knowledge, and express feelings. In its written form it is used to commit laws and contracts to paper that determine and regulate our lives. That there is more to language than just signs and rules for their combination is illustrated by the fact that some utterances can be seen as acts [Austin, 1975] due to their effects in the world. Our language determines many aspects of our identity, and is sometimes even attributed magical powers, e.g., in taboo words [Crystal, 2003]. By some, language is seen as a means of thinking designed to facilitate human cognitive processes, and its use for communication as a mere byproduct [Chomsky, 1993]. This illustrates that language is deeply linked to human cognition and thought. And we still know little about the exact nature of this link. This makes automatic analysis of language particularly challenging.

In this thesis, we focus on a confined area of language use, namely written texts in the scientific domain. We aim to facilitate and expedite exploration processes that help humans view and explore large bodies of documents. Thus knowledge about the information contained in such collections can be attained. This focus arguably removes some levels of complexity, such as dealing with audio signals and phonetics and phonology [Clark et al., 2007] or high noise levels that texts in other domains exhibit [Han and Baldwin, 2011]. However, the general challenge that language is highly ambiguous and context-dependent, remains. Despite steady advances in natural language processing (NLP), high-accuracy automatic analysis

and summarization is not possible due to the complexities on multiple levels. And there is reason to believe that this is not to change any time soon [Searle, 1980].

In this work, we support human users with the exploration of text document collections. This is achieved through the design and development of interactive visual interfaces that support cognitive processes based on the visual capabilities of humans. The young research field of Visual Analytics (VA) [Wong and Thomas, 2004] has been devised to research such interactive methods. It is rooted in Information Visualization (InfoVis) [Card et al., 1999] and aims to support human reasoning and sensemaking processes [Ribarsky et al., 2009; Keim et al., 2008, 2010]. This coupling between machines and humans makes VA highly suitable to support users with text exploration and analysis tasks. It allows to integrate highly efficient data manipulation and pattern recognition capabilities of computers with world knowledge, decision making, and linguistic skills of human users. This way, automatic methods can be used to abstract, aggregate, and characterize natural language texts to create visual representations. Users, in turn, can work with these either directly to gain knowledge, or as navigation aids that efficiently guide the way to relevant information. In addition, users can steer and adapt automatic methods through interaction to adapt the results to their interests or analysis needs. This thesis provides novel approaches that integrate humans and computers to offer support with exploration tasks on large text corpora. These approaches help users to organize and extract knowledge from these corpora without having to read each single document in detail. For this, we extend the state-of-the-art in VA in several ways. Throughout this thesis, we motivate and discuss each of our contributions.

1.1 Motivation

Most of the documents published today are produced, and to a large part also distributed, digitally. One benefit of this technological revolution is the ubiquity of information through easier distribution and access. And the availability of affordable mass market devices such as tablet computers and e-book readers not only allows us to carry huge amounts of documents with us at any time. They also facilitate access to new information from everywhere. However, fast and ubiquitous access to information is just one side of the coin. Almost instant access to a large number of texts allows us to create large corpora from which is possible to generate new insights. Documents cannot only be searched to attain information that exists within them, but their contents can be correlated and combined to essentially derive new knowledge. In the realm of scientific literature this helps to tackle long existing challenges, such as related work search in new disciplines, trend analysis, and

evaluating important authors. Even the generation of new scientific insight seems to be possible, as demonstrated by Swanson [1991] and Swanson and Smalheiser [1997]. The only way to fully leverage these opportunities is by connecting humans with the data crunching abilities of computers. The reason for this is twofold. Firstly, fully automatic processing is not possible as discussed previously. Thus, although automatic methods can speed up analysis and help finding interesting patterns in text data, human world knowledge and intuition is always needed for correct interpretation. Here, VA methods foster the integration of humans and algorithms, and provide visual schemes for conveying provenance and uncertainty information. Both are important to gauge the capabilities of automatic methods, identify potential mistakes, and, in the end, build trust into their results. Secondly, intentionality is inherently human, and users are the ones to explore data sets, interpret findings, and decide what to look into in more detail. It is thus important to let users steer and interact with automatic methods to adapt and configure them for the analysis questions they are interested in.

We focus particularly on the early stages of analysis, during which exploration plays an important part. As Tukey puts it: "Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider." [Tukey, 1977]. This is true for two aspects of exploration that can stimulate the generation of questions from data. Often, large data sets are collected and their contents are either largely unknown, or knowledge about it is too shallow or cursory to be able to look into specific aspects of the data set. An example would be large amounts of micro blog messages that are collected world wide in a certain time frame. What type of topics are contained within the data set? Is there any unusual correlation of these topics to other message properties, such as the location they were sent from, or the language they are written in? Are there any other conspicuous properties in the data that might be worth looking into? Such questions are answered during exploratory analysis. Alternatively, the data set has been assembled according to clearly defined aspects with respect to its contents, but the analysis question is only vaguely defined. In this case, users are enabled to view the data and find clues about what might be worth exploring to substantiate their analysis questions. An example for this is related work search in a yet unknown discipline, in which users first have to learn about the conventions, the language, and how much work interesting to them exists in that community. During exploration, new potentially relevant artifacts can always turn up. This can include, e.g., frequent citations to another discipline that contains a whole new set of additionally relevant work.

For scientific document collections, free exploration is particularly beneficial. The reason for this is that they are rich in metadata and thus provide many different data dimensions that can be correlated. Author lists and connections between

these single authors, temporal dynamics, and in particular citations, are important aspects that embed a document into its scientific context and relate its content to that of other documents. Here, VA can help to visually connect these data aspects, and let users grasp and understand connections within the data. In addition, of course, there is also a practical motivation behind our work. As researchers we work and analyze scientific literature daily. We are thus enthusiastic about the improvements and progress with which exploratory visual text analysis is able to impact and improve our daily work.

1.2 Research Questions

Based on these goals and the state-of-the-art in exploratory approaches for text and scientific literature data, the following research questions have emerged in the course of this thesis:

Question 1 *How can exploratory analysis of text data be supported on arbitrary levels of granularity in a flexible and configurable fashion that keep users close to the actual data?*

Question 2 *How can automatic classification methods be trained effectively in a way that fosters information exchange between humans and computers to store the insights gained during exploratory analysis of text corpora?*

Question 3 *How can NLP experts be supported with interactive visualization to analyze and debug their text processing methods in order to improve automatic text mining facilities of VA approaches?*

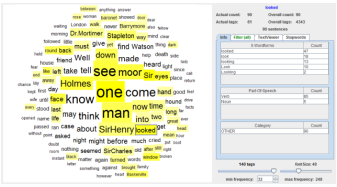
Question 4 *How can the contents of document sets be visually integrated within their intellectual context based on available metadata information in a way that helps create new insights into scientific communities?*

Throughout this thesis we present and discuss multiple text exploration and analysis approaches that each tackle at least one of these questions.

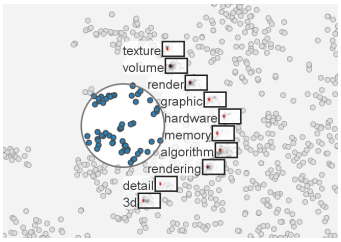
1.3 Contribution and Structure

Our contribution consists of five main approaches that each address one or more of the above research objectives. We list them in the sequence they appear throughout this thesis and include a brief description of how they fit into the theme of this

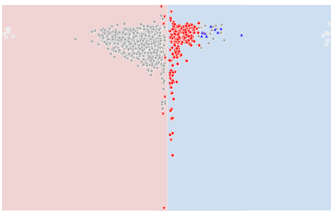
work. An additional part of the contribution of this thesis is the development of a classification scheme for text VA approaches. It classifies approaches based on abstract types of visual knowledge generation tasks, the type of integration of machine and users, and the linguistic level that text data is processed on (see Chapter 8).



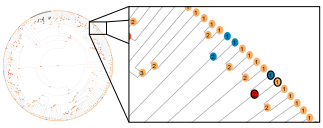
The Word Cloud Explorer is an approach that extends traditional word cloud visualizations [Viégas and Wattenberg, 2008] to a full-fledged text exploration and analysis framework. It provides a main view that consists of frequency term clouds from one coherent body of text. Users can interact with this view to get additional information about word occurrences and co-occurrences. In addition, term and term category filters allow users to constrain shown terms by co-occurrences or term type and thus solve many types of advanced analysis and exploration tasks. The approach exemplifies the integration of automatic text processing and human users through straightforward visual metaphors that make it versatile and well-suited even for casual users [Bosch, 2014].



DocuCompass is a technique that allows for flexible exploration of document spatializations. Moving from single bodies of text, as supported by the Word Cloud Explorer, to collections of large numbers of documents, this approach visually retains document boundaries to structure a text data set. The approach facilitates the exploration of document spatializations on arbitrary levels of granularity, freely chosen by the user. It thus constitutes a highly flexible method for exploration and is compatible with every possible type of text spatialization. It can be configured using different types of document characterizations, filtering and interaction facilities.



Text Classification can help to organize and filter large amounts of text data at high speed. In VA, classification has been used to filter and identify many different types of data instances. This helps users manage large amounts of streaming data to concentrate on the aspects most important to them during exploration and analysis. Our approach to text classification is designed to interactively train a classifier by exploring a document space and finding good instances to label. These are then used to create a classifier whose state is visually represented and can be explored by the users, who in turn label new examples to improve its performance iteratively. This is an example approach in which users and machines learn from each other during exploration by exchanging useful knowledge that is beneficial for the analysis on both sides. The resulting classifiers are useful tools for future retrieval of documents and can be applied for data foraging in text VA approaches.



FeatureForge is an approach designed to debug features and learn about their benefits and drawbacks. It helps NLP experts to design new features that improve machine learning methods on language data. Similar to the previous approach it gives insight into the state and performance of classifiers, but focuses specifically on the improvement of the set of features that describe single data instances. The previous three approaches are designed for users who are experts for the data set to analyze. FeatureForge, on the other hand, is designed for linguists that design and implement automatic methods to process language data. It helps them to test and improve the data processing methods that underlay text VA approaches. FeatureForge illustrates how VA approaches for data analysis can be coupled on multiple levels, linguistic data and domain data in this case. Here, the knowledge gained during the analysis of the linguistic data helps with the exploration on the domain level.

CiteRivers specifically focuses on the exploration of collections of scientific articles. Until recently, such methods have either supported text exploration or visual analysis of community structures based on metadata or citations of scientific documents. Arguably, both aspects are important, however, previous approaches look at them separately. CiteRivers, in contrast, facilitates their joint exploration. This gives users a different perspective on a document set during exploration, and lets them learn, e.g., about the connection of a field to others. It thus provides deeper insight through correlation of different data aspects. In addition, the approach provides multiple interactive views on the scientific literature data that allows for the analysis of prominent topics and trends, important authors within a community, and highly cited and trendy publications. In comparison to previous approaches, it helps users gain a more holistic view on collections of scientific articles.

These five approaches constitute the bulk of this thesis' contribution. In addition, further collaborative projects have been pursued during the work on this thesis. These collaborations resulted in additional publications that bear relevance to various degrees to the described topics. We briefly mention them in the following, and refer to them throughout this thesis wherever they are relevant to our discussions.

Digital humanities is an emerging, interdisciplinary field that aims to support literary and social scholars with interactive visual methods. Here, we contributed to the creation of a focus+context approach for navigating through single [John et al., 2015] and multiple texts for comparative analyses [John et al., 2014]. Furthermore, we were able to bring in our knowledge about interactive classifier creation to the development of a search engine for environmental conditions [Vrochidis et al., 2012], to social media analysis for emergency management [Bosch et al., 2013], and to predicting movie box office success [Krüger et al., 2013]. We further contributed to an interactive visual approach to conference management that helps to predict visitor streams and room occupancies [Krüger et al., 2015]. Beyond VA approaches for text analysis, we were able to contribute with our methods to approaches for eye tracking data on video stimuli [Kurzhals et al., 2014, 2015], and for the visual analysis of movie contents [Kurzhals et al., 2016]. In addition, we were involved in creating and publishing a literature survey about interactive visual approaches to patent and scientific literature analysis [Federico et al., 2016]. Finally, we have

been working on and constantly updating a data set of publications from the IEEE VIS(Week) during the work on our approaches. We have used this data at various stages for use cases and user feedback sessions in most of our projects. Whenever this data set appears in this thesis, we describe its state at the time we were conducting the user feedback session. We are particularly proud to have been able to contribute parts of our data and knowledge about data cleaning and extraction gained during the project to a comprehensive and constantly updated set of VIS publications. This data is now available online as a test and benchmark data set for the entire community [Isenberg et al., 2016]. We hope to contribute with this to the future development of visual interactive approaches to scientific literature.

CHAPTER 2

Foundations and Background

This chapter discusses relevant background and work related to the topic of this thesis. It starts with an introduction to Visual Analytics (VA) focusing on sense-making and exploratory analysis. Then, we discuss relevant background on natural language processing (NLP) and machine learning (ML), focusing on the techniques that are used throughout the chapters of this thesis. Finally, relevant related visual text analysis and scientific literature analysis approaches are reviewed and differences to our methods are discussed.

2.1 Visual Analytics

Visual Analytics [Wong and Thomas, 2004; Thomas and Cook, 2005; Keim et al., 2006, 2008, 2010] is a young research discipline that is concerned with helping humans generate knowledge, make new discoveries, and produce insights about data and the world in general. Its primary goal is to exploit the capabilities of the human visual system as a high-bandwidth connection to the brain and its visual pattern recognition abilities [Ware, 2012] to connect both computers and humans. Visualization, here, serves as a means of communication that facilitates information exchange between both sides. Interaction with this visualization is then the language of humans to communicate back to computers. Such an integration of humans and computers is crucial to create the essential tools for a world in which information is the primary resource. Knowledge and insights into the world and our surroundings have always been at the center of human pursuit, and are gradually taking an evermore prominent place in our modern society. VA essentially combines

the data crunching abilities of machines and intuition and world knowledge of humans to organize, analyze, and gain insights from large amounts of data. In many of today's applications of VA, the insights gained have immediate practical applicability, e.g., for human decision making. However, all domains in which the acquisition of human knowledge can be based on large data sets are potential fields of application for VA. An example are the nascent field of digital humanities that uses VA methods to help, e.g., literary scholars with their knowledge gaining process from literature [Koch et al., 2014; John et al., 2014].

As a young and vibrant research field, the practical applications of many current VA approaches are an answer to today's opportunities that come with the ubiquity of computing devices. As a consequence, more and more data is being recorded and becomes available of increasing number of aspects of our lives and at a growing pace. This leads to huge, often heterogeneous, incomplete, noisy, and constantly updated data sources that hide valuable knowledge within them. While fully automatic analysis is an option to harness this information, there are circumstances that render it either unfeasible or prohibit it altogether. While automatic algorithms are capable of analyzing patterns within a data set, sorting data according to similarities, or uncovering dependencies between data variables, only human users can assign meaning and interpretations to such findings. Humans in the loop are thus essential in order to generate knowledge about data, and thus to tackle and solve complex analysis tasks.

Insights and knowledge gained through visual interactive methods can be used for further analysis of the same data set by steering automatic processes and adapting and refining analysis goals and strategies. Scott et al. [2002] corroborate this by showing how human understanding of computational problems and corresponding human feedback can help algorithms find better solutions faster, and at the same time help humans better understand the problem and increase their trust in the solution. Analysis circumstances in which this is particularly helpful are typically a combination of the following three. Firstly, the contents of a data set are either entirely or partially unknown. This means that users have to learn about the contents of the data set using exploratory data analysis methods. Based on the gained knowledge, users learn about what information can be acquired from the data and which problems can be solved with it. The approaches within this thesis mainly focus on this scenario, but they also partially support the following two. Secondly, the users have entirely or partially unknown analysis goals. An example for this is related work search, where users need to explore a document set to find documents related to their own current research project. The terms and formulations used to describe such work are then discovered and refined, which leads to incrementally increased coverage of the analysis. Koch et al. [2011a] present an example for such a scenario that helps iteratively refine queries for

patent retrieval through interactive visualization of query result sets. Thirdly, VA can help with the analysis of complex or particularly noisy or faulty data. In such a scenario, users are supported in gauging the quality of the data, analyze and learn from imperfect data, or correct faulty analyses to steer automatic processes to useful results. An example for this is presented by Maciejewski et al. [2008] for the analysis of temporal and spatial health symptom data to detect disease outbreaks.

VA is still a very young discipline, with the term's first appearance in the title of an editorial by Wong and Thomas [2004]. It defines VA as an "approach to combine the art of human intuition and the science of mathematical deduction to directly perceive patterns and derive knowledge from them" [Wong and Thomas, 2004]. VA combines information visualization with many other disciplines, including cognitive science and statistics to create a new science of reasoning that helps view, understand, and learn from large amounts of data. While information visualization focuses on the visual encoding of data, VA combines such encodings with interactive filters and feedback mechanisms that foster human cognitive processes. Shortly after this, a thorough definition and road map of the field of VA was published by Thomas and Cook [2005]. In the wake of the attacks on the World Trade Center in 2001, they position VA as a research field that can help to eliminate future threats through the analysis of massive intelligence data. Besides its political stance and agenda, it comprehensively lists technical and scientific challenges and goals for this new field. For the analysis of text documents, Thomas and Cook [2006, recommendations 4.2-4.5] recommend multiple research directions, some of which we take up in the research presented in this thesis. Their recommendations include the enhancement of simple token-based processing methods and visualization with additional information, such as named entities or multi-words. This is a central feature of our approach in Chapter 3. They also recommend the development of real time characterization methods for documents in streaming scenarios, as we do with the approach in Chapter 5. Furthermore, providing additional information for the processing and analysis of text data is listed as a goal. We do this in Chapter 7 for scientific literature. In addition, the approach in Chapter 6 aims at supporting the development of automated methods that incorporate additional linguistic information from the texts.

Interacting with data and its visual representations is crucial to help human cognition. The findings of Kirsh and Maglio [1994], that suggest that action can be an important tool to help cognition, stresses the importance of direct interaction with visualizations to help understand and evaluate data. In contrast to Thomas and Cook's groundwork for VA, Keim et al. [2008] stress its third component, automatic data mining and analysis. They developed a prototypical pipeline to describe the VA process, which is depicted in Figure 2.1. The pipeline starts from

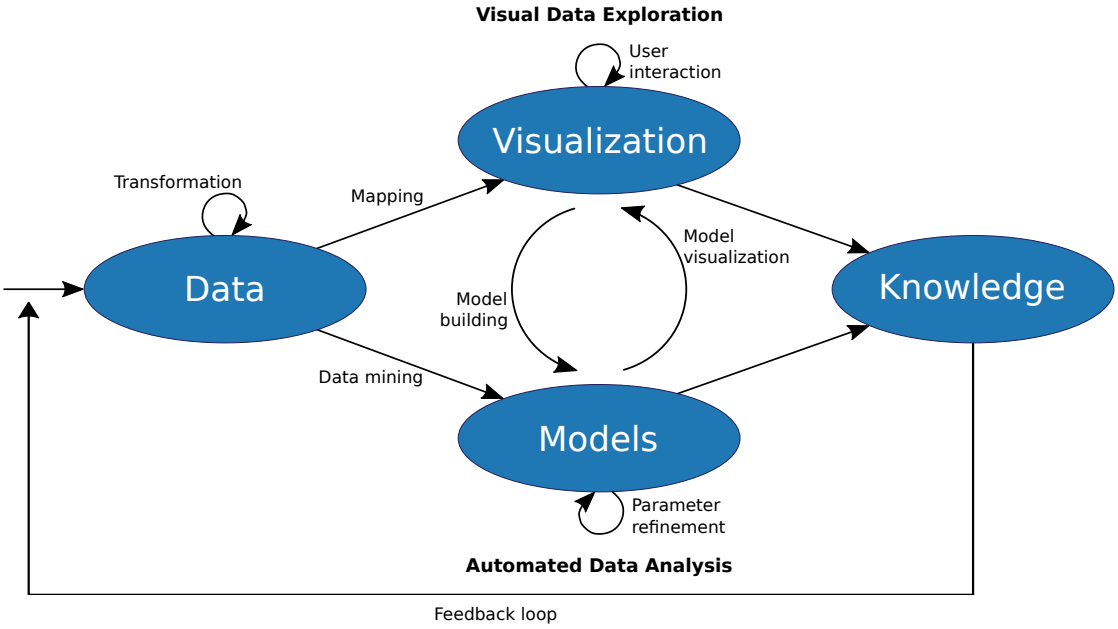


Figure 2.1 — Keim et al. [2008]'s pipeline that models VA as a process that starts from raw data visualized and consumed by automatic model building algorithms. Users can then explore and analyze the data as well as the models because both are represented visually. Influencing data and models is also possible through directly interacting with them. This process helps users to gain insight into the data and allows them to ultimately derive knowledge from it.

raw data that is visualized and consumed by automatic model building algorithms. As we are mostly concerned with the analysis of natural language texts, in our case, the data is mostly textual data, but can also be authors, citations, or any other type of meta data of scientific documents. The models are created automatically from the data by identifying patterns and correlations within the data through the knowledge discovery in databases (KDD) process [Fayyad et al., 1996]. Again, most of the models we are concerned with are those of natural language processing algorithms. A VA approach lets users explore and analyze both data and models, providing feedback mechanisms that allow for operations on both. With this interaction loop, users are able to steer the refinement of the models according to their analysis needs. This helps them to understand the data, and, through human cognition and sensemaking processes, to ultimately derive knowledge and insights [North, 2006] from it. However, research on how insights are gained through cognitive processes to guide design decision for visualization and interaction methods has been scarce until now. This is pointed out by several publications that reflect on the state of

VA research [Ribarsky et al., 2009; Thomas and Kielman, 2009; May et al., 2010], mentioning it as future challenges of the field at which pragmatic solutions have been targeted [Shrinivasan and van Wijk, 2008]. On the application side, however, VA has seen huge success including the inception of a corresponding conference in 2006. This has jumpstarted research on new methods and approaches, with much fundamental work on important technical issues such as different aspects of scalability, and many effective approaches to solve data analysis problems in a wide variety of domains. Meanwhile, it has become an integral part of call for papers for all major visualization conferences. For insights on the VAST conference and its relation to the other two VIS conferences see the use case in Chapter 7 or Heimerl et al. [2016a].

2.1.1 Data Visualization

VA is rooted in visualization, a discipline that investigates visual representations to convey information to humans through graphical displays [Ware, 2012]. These representations aim to facilitate human cognitive functions [Liu and Stasko, 2010] including the processing and the correlation of information [Larkin and Simon, 1987; Scaife and Rogers, 1996]. Visualization has evolved into two different fields, Information Visualization (InfoVis), and Scientific Visualization (SciVis). The latter is concerned with the visual depiction of data that is grounded in our physical world. It may come from measurements of physical quantities or processes, such as particle flows or magnetic fields within a 3-dimensional space. Other examples are astronomical images or particle data that capture chemical processes. SciVis approaches are designed to help science researchers analyze and understand their data and the output of their simulations. InfoVis, on the other hand, aims at mapping abstract data to visual representations to support users with its analysis [Fekete et al., 2008]. Examples for such data are natural language texts, computer network data, financial data, and social network data. InfoVis and SciVis, although they work with different data sets, have similar goals and share a large number of common techniques and methods. Both are governed by the perceptual properties of the human visual cortex and human cognitive functions. They are consequently governed by much of the same design guidelines and principles [Ware, 2010; Tufte, 1990].

As this thesis is concerned with the visualization and analysis of natural language texts, we will henceforth concentrate on the InfoVis side of things. InfoVis has been defined by Card et al. [1999] as "The use of computer-supported, interactive, visual representations of abstract data to amplify cognition." Their information visualization pipeline (see Figure 2.2) describes the underlying process of data transformation for InfoVis. Although alternatives and different versions have

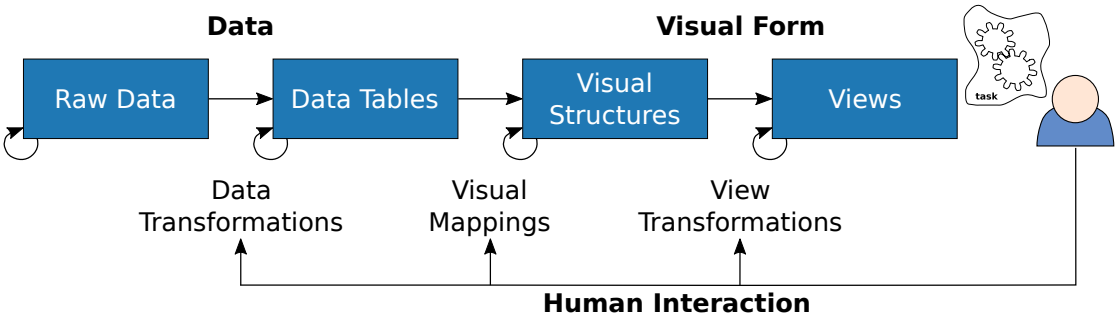


Figure 2.2 — Card et al. [1999]'s information visualization pipeline.

been discussed, e.g., by Chi and Riedl [1998], we present it here as the most widely used one in the visualization literature. The central data structure of the pipeline are the data tables in which data is stored after being transformed from its raw form. This can be, e.g., a novel from which literary characters and their relationships are being extracted and stored in tables. The resulting tables would then contain links between persons as data items including a specification of the type of their relationship. Assuming we want to visualize this information as a graph, it would then be transformed into a table containing all visual elements comprising nodes that represent persons and edges that represent relationships. These visual structures can also be implemented as tables, as they are in *prefuse* library [Heer et al., 2005], an InfoVis library that is designed according to this pipeline. The view transformation transforms these visual elements into graphical primitives that determine the final images drawn.

In addition, the model supports interaction at each stage of the transformations to allow for user intervention. Interaction, here, is an important tool for users to support the knowledge discovery and insight generation process [Pike et al., 2009]. The crucial step for visualization of the data are the visual mappings. They determine the type of visualizations used for the data at hand, and the previous and subsequent transformation steps by and large depend on it. Shneiderman [1996] defines a task-by-type taxonomy for different visualization types and provides his now famous mantra as a guide for the design of interaction and data filters "Overview first, zoom and filter, then details on demand." For some data types, the initial transformations from raw data to data tables are mostly trivial, especially for data that is already in the form of tables, e.g., computer network data or financial transactions data. For other types of raw data, such as images, video, and text, these transformations can be based on the results of entire research disciplines that provide the techniques to extract structured information from them, e.g., computer vision, or computational linguistics. For language and text data, Collins [2010] identifies a "linguistic visualization divide" in that he argues

that these data transformations are often not done with NLP methods that are sophisticated enough to support the analysis as well as possible. This InfoVis pipeline, of course, is also relevant for VA of abstract data such as text, since InfoVis plays an important part within VA. Koch [2012] shows how automatic data mining can be included within this pipeline by integrating the VA process from Figure 2.1 and the InfoVis pipeline from Figure 2.2 for VA approaches.

2.1.2 Insight Generation through Exploration

Effective interactive visualizations are important to accurately convey information within a data set to human users. However, as Card et al. [1999] put it, "The purpose of visualization is insights, not pictures.", thus defining human knowledge as the ultimate goal of visualization approaches. Van Wijk [2005] provide a model that quantifies the cost of human knowledge generation through interactive visualization. The definitions of what exactly insights are differ within the visualization community, although attempts have been made at categorizing them [North, 2006]. Based on a review of visualization literature, Yi et al. [2008] take a procedural stance and identify four processes that lead to insights in visualization approaches. These are *provide overview*, that lets users learn about the contents of a large heterogeneous data set, or its yet uncharted areas. The *adjust* procedure helps to gain insights by enabling them to adjust the level of abstraction of their view of a data set, while the *detect pattern* procedure relies on the identification of patterns in the data, such as recurring sequences or outliers. Finally, *match mental model* produces insights by providing artifacts that users can interpret by mapping them to artifacts of their world knowledge.

These processes are all relevant to the approaches contained within this thesis, which aim to support the early, exploratory stages of analysis during which knowledge about the concrete analysis questions is generated and information needs [Hearst, 2009] form. According to Stasko [2014], this knowledge not only provides insights for subsequent analysis task, but also an "overall essence" of the data, and "confidence [...] and trust" in the data. It is thus one of the reasons why visualization provides value for data analysis. From our experience, when users are confronted with a novel data set with unknown contents, they typically pass through two stages during exploration. These stages are passed depending on the information need of the user and her posterior knowledge about the format and contents of the data set. We call these phases the *collection exploration* phase, and the *information exploration* phase. Imagine a historian that wants to learn about the history of chemistry by analyzing a large digitized set of issues from a time span of 50 years of multiple chemistry journals. With an initial idea about the data set, she wants to learn more through exploration. She thus loads the data set into a VA system that

supports the collection exploration phase during which questions about the data set are at the center of analysis. In other words, this phase of the analysis answers the question "What is the scope and quality of the information contained in the data set, and what type of questions can it help me answer?". All of the processes for insight generation listed above can be employed to generate this knowledge. The questions to be answered comprise:

- *Breadth of information*
  What is the scope of the data set in terms of topics and information that is present, and is it complete or are parts of it missing? For example, the historian learns that the data set consists of five journals that all have differing topic profiles and focus on diverging subfields of chemistry. She also finds that the data set covers the same 50 year period for each of the journals, but for one of them, some issues seem to be missing. The data set is thus not entirely complete. In addition, she learns that the data set contains ample meta data for each of the articles, including authors and their affiliations, and complete lists of references seems to be available, too.

- *Depth of information*
  How deep is the information that is present within the data set? Examples of insights about the depth of information in the data set for our user are that all five journals have different standards concerning the depth of the related work section of the papers. Articles from some journals tend to provide longer discussions of related work than others. In addition, she discovers that two of the journals started publishing extended abstracts from a particular conference. Although these avidly describe new research projects and ideas, they are not complete in a way that allows for repeatability of the research described. They also do not contain a lot of information on results. She decides to be careful with handling these types of publications and with integrating them into her later analyses.

  The previous two aspects are about what kind of information is contained within the data set, while the following two deal with how this information is represented.

- *Information encoding*
  The user, for example, finds out about the conventions for author names in the different layouts. She also learns that they have changed over time for each of the journals. This helps her to interactively create a process to automatically extract and visualize author information from the documents. It also helps with extending her mental model of the data set for easier navigation and to decide about what the information from the data set can

be used for. Further examples in this category are, e.g., chemical compounds contained within tables, or even file naming conventions or data base schemes of the data set.

- *Language*
  The type of language used is another property of text data sets that is important to know before analyzing its contents in more depth. This can be exploring the level of language used and the terminology that is used to describe facts in order to gauge who is the addressee of the texts. Scientific articles, of course, are written in a technical and impersonal voice, but the user realizes during initial exploration, that one of the journals obviously publishes articles in English and French. This means that if she wants to analyze these articles in more depth, she will need translations.

After this initial exploration phase, the second phase typically starts quickly, and users start to explore the actual contents of the data. Both phases can run in parallel then, with the user switching back and forth between both. In the information exploration phase, the user explores the contents of the data set that catches her attention. This depends on the salience of data instances within the visualization, but also the interest and background knowledge of the user. Although this makes it particularly difficult to identify abstract categories for the insights that play an important role in this stage, we found that three types of data instances, or groups and sequences thereof, are of relevance. Firstly, the identification of common and often recurring data items is important to learn about the general contents of the data set. For our historian, these can be, e.g., articles that exhibit certain popular patterns or that are concerned with a specific topic. Multiple groups of popular topics could also be uncovered using clustering techniques. Here, the user can identify interesting data artifacts, such as authors she often comes across during exploration, or topics that are particularly popular within certain journals. Secondly, in contrast to the identification of frequent data patterns, outliers are also highly relevant. This could be extraordinarily highly cited papers, or authors that are particularly productive and have many co-authors. Thirdly, identifying patterns is also of interest in this stage. These can be temporal patterns, such as topics that become more popular over the years, or groups of authors that tend to work closely together on their publications. During this stage, it is important to give users freedom in exploring whatever catches their attention by supporting the above listed insight generation processes as comprehensively as possible. This fosters serendipitous discoveries and allows users to create and maintain a mental map or sketch of the data set. Exploration during this stage has some similarities to the process of berrypicking [Bates, 1989] in information retrieval, and helps users to exploit the data for more targeted analyses later [Tukey, 1977].

2.1.3 Sensemaking

To generate such insights, a cognitive process called sensemaking plays an important part [Yi et al., 2008]. There are gradually differing views in the literature of what sensemaking is, and how it can generate knowledge. We will briefly mention two of the theories, as they are the ones that are able to provide some theoretical background to knowledge generation during visual interactive data exploration. The most prominent theory in the visualization literature is the one by Pirolli and Card [2005]. It is based on the notion that the core of sensemaking is a "learning loop complex" through which humans generate knowledge by continuously creating and adapting schemas and revising them based on the data they observe until all relevant data points are in agreement with it [Russell et al., 1993]. They propose a waterfall model of sensemaking that is derived from the typical procedure of an intelligence analyst. The model is depicted in Figure 2.3. It consists of several

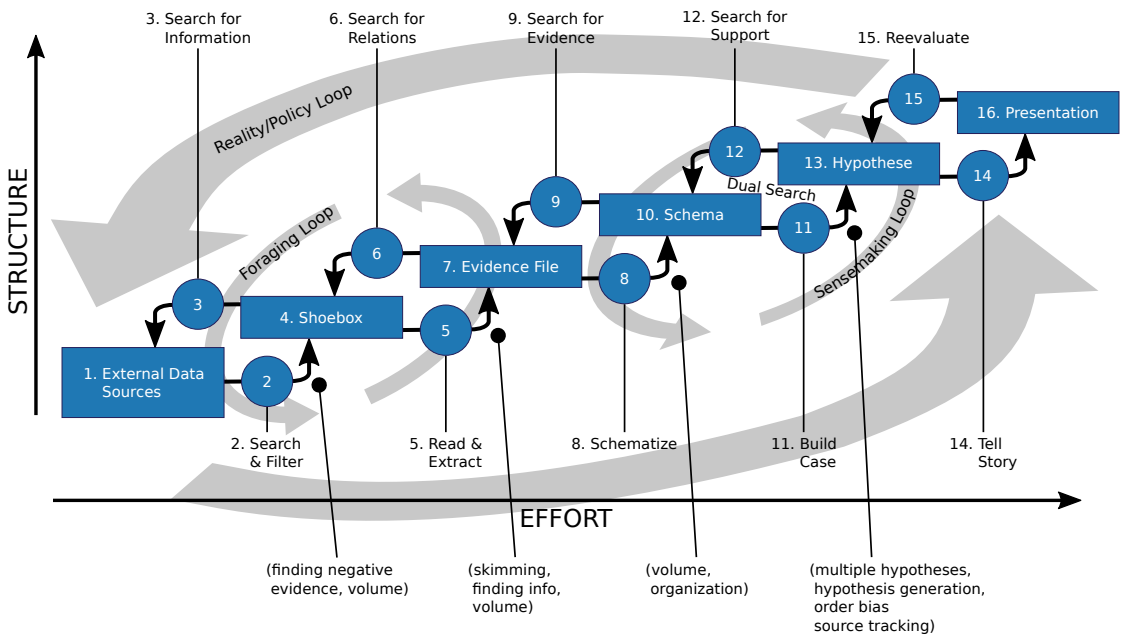


Figure 2.3 — The waterfall sensemaking model according to Pirolli and Card [2005]. It consists of two loops, the *foraging loop*, and the *sensemaking loop* that both play a role in exploratory analysis. Roughly, in the former loop, information is collected from data, while in the latter, collected information is correlated and organized to form a coherent story.

stages that are all linked to each other in a sequence. Each of them is associated with one of two loops, the *foraging loop* during which relevant data is collected, and the *sensemaking loop*, during which the data is evaluated and a coherent story

is constructed from it for final presentation. The stages for the foraging loop are *external data sources*, which is the source data repository on which the analysis is based, the *shoebox* that stores relevant data items that have been identified, and the *evidence file* in which the relevant information extracted from the data are collected. The foraging loop is thus the loop that is responsible for identifying, collecting, and stripping pieces of information from the data sources. The stages of the sensemaking loop comprise the *schema* that is created to organize and coalesce evidence into larger chunks of coherent pieces of information, the *hypotheses* that integrate all of the evidence chunks into a coherent story that can be updated, refuted, or confirmed by further information, and the *presentation* as the final stage to communicate the results, e.g., to peers or superiors. It is important to note that when developing VA approaches, according to Pirolli and Card [2005], some of the stages can also happen in the analysts' mind and do not necessarily have to be supported explicitly, depending on the analysis task to be facilitated by the approach. The process that the model describes can be either executed in a bottom-up (steps 2., 5., 8., 11., 14. in Figure 2.3), or a top-down (steps 15., 12., 9., 6., 3. in Figure 2.3) fashion. Bottom-up transitions abstract further away from the data towards derived knowledge or insights, while the top-down transitions search for further pieces of information corroborating or refuting derived knowledge or insights. Users can freely invoke each of the steps depending on their analysis needs.

In their *technology leveraging points* for the foraging loop, Pirolli and Card [2005] describe the "exploration-enrichment-exploitation tradeoff", which is the process of consecutively narrowing down on the relevant data items by generating increasingly precise retrieval specifications based on increased knowledge about the data set. This is a process that we support with the approach presented in Chapter 5. In addition, the approach from Chapter 4 is designed to generated a very low overhead when shifting attentional control to various granularity levels of exploration. This is another leverage point for the foraging loop [Pirolli and Card, 2005]. During the exploration phase of data analysis, both the foraging loop and the sensemaking loop play an important role, as the identification of interesting or salient data items as well as their interpretation are important to get insights about the information contained within a data set.

However, this model assumes very clear-cut analysis goals and a starting and end point that typically do not exist during free exploration. Sacha et al. [2014, 2016] present an alternative, integrated knowledge generation model for VA. It combines the VA model from Figure 2.1 with a model of human cognitive processes during interactions that consists of three loops. These build on one another and are called from bottom to top: the *exploration loop*, the *verification loop*, and the *knowledge generation loop*. The exploration loop comprises free exploration of the

data set in order to generate analysis questions from it. The verification loop uses the facilities of the exploration loop to guide it towards findings that help create, confirm, modify, or refute hypotheses. The knowledge generation loop derives new insights from the constant creation, confirmation, modification, and refutation of hypothesis. Newly gained knowledge also serves as fertile ground for the creation of new hypotheses by the user. While this classification of cognitive functions includes exploration as the basis of knowledge generation, it leaves open the question about the nature of the cognitive processes that underlie visual information exploration and analysis.

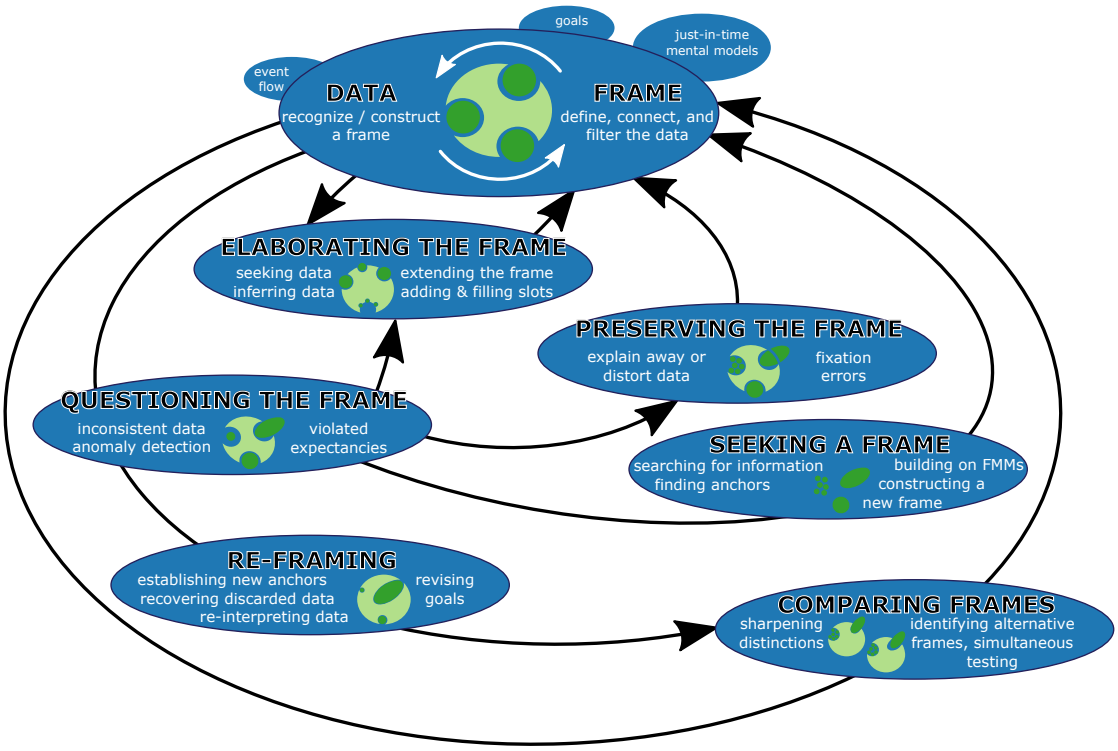


Figure 2.4 — According to Klein et al. [2007] sensemaking is a complex, iterative mental process that can take on seven different forms.

The data-frame theory [Klein et al., 2006b,a, 2007] has some answers to this. It views sensemaking as a constantly ongoing cognitive process to explain the world around us. The data-frame theory is based on a notion of sensemaking for individuals and groups [Weick, 1995; Weick et al., 2005], e.g., a group of firefighters that have to function in and make sense of situations collectively to safely handle tough, or even life-threatening events [Weick, 1993]. Some VA environments also support sensemaking as a social and collaborative process [Heer and Agrawala, 2008]. The theory according to Klein et al. [2007] assumes the two elements *data*,

and *frame*. Data are sensations that reach a human's mind, while frames are logical constructs or stories that human cognition builds around the incoming data. Data points that are sufficient for the creation of a frame are called its *anchors*. For each new situation or event, our minds constantly construct frames for incoming data that try to explain it, even if we are not conscious of it. This means that incoming data points are constantly matched against an existing frame. If possible, they are integrated into the story line of the current frame. Otherwise, the frame is rejected and a new one that is consistent with the new data point is created. Once the situation can be explained satisfactorily and no new conflicting data comes in, the sensemaking process for a specific situation loses its appeal and may eventually cease entirely. Sensemaking, according to Klein et al. [2007] is a complex process that can take on at least seven different forms. We have reproduced the diagrammatic depiction of the sensemaking loops in Figure 2.4. It shows that this is not a tidy intellectual process that is executed after careful deliberation or even fully intentionally. Users may switch from a conscious to an unconscious process depending on how unusual the situation is, and whether a matching frame is readily available. Also, users often rely on just-in-time models [Klein et al., 2007], i.e., incomplete models about processes they have no exact knowledge about, e.g., how a plane functions. According to Klein et al. [2006b], several psychological concepts are closely related to what we call sensemaking. These are all human cognitive hallmarks including creativity, curiosity, comprehension, mental modeling, situation awareness.

Figure 2.4 shows the seven different forms of sensemaking stipulated by Klein et al. [2007]. The sensemaking loop, in any of these forms, can start after new or disrupting information comes in. In the VA context, this happens when users come across surprising, contradicting, or unexpected information artifacts. New frames for data are constructed with influence from just-in-time models, the user's goals, and event flows users have in mind. The repository of frames that users know is the world knowledge and experience that they bring into the analysis. According to Klein et al. [2007], experts and non-experts have the same abilities for sensemaking, only differing sets of frames. Sticking to a frame to explain data generates expectations and as a result filters and may even construct data. Thus different users are not necessarily aware of the same data points during analysis, depending on the frame and thus their background knowledge and experience. Besides matching a new frame to selected anchor points in new, incoming data, the forms of sensemaking are: *elaborating a frame* in which a person actively looks for or derives new data to add more detailed information to a frame; *questioning a frame* happens when data conflicting with the current frame emerges or expectancies that a frame ensues are not met; *preserving the frame* is the decision of users to keep a frame which may lead to distorting or ignoring data. Depending on whether the

ignored or distorted data is actually unreliable this can be a beneficial decision; *seeking a frame* is the active process of looking for data as anchors to a find a new frame after a previous one has been dropped; *re-framing* in contrast to seeking a frame involves revisiting old potentially discarded data and looking at it in a new light, goals can also be redefined or rephrased during this process to easier frame a situation; *comparing frames* is tracking of up to three different frames in parallel for some amount of time. They explain the same data and users to compare them and resolve particularly complicated situations [Klein et al., 2007].

Based on the reasonable assumption that sensemaking happens during data analysis the same way as it takes place in other situations, this model can be used as a basis for the design of interactive data analysis approaches. In fact, Klein et al. [2007] claim that this model is indeed applicable to interactive software. We have compiled a few points that we took away from this work in sensemaking and where they have inspired the design of our approaches. Anchors or data points that integrate into frames easily are fundamental for many forms of sensemaking. From this, we can derive the importance of access to and overview of a large number of diverse data instances in VA approaches. This is supported particularly well by the approaches from Chapters 3 and 4. Starting from an overview, it allows users to freely explore document data, filter, and connect data points according to different aspects. Most importantly, it allows users to freely adjust the granularity of exploration to minimize the costs of switching between different levels and contexts. This allows users to find links between data instances faster to support or question frames. Linking of different data aspects is also one of the main goals of the CiteRivers approach (Chapter 7) that focuses on a visual integration of citations and the contents of scientific publications, but also supports other aspects. Through visually linking them, it facilitates the informed identification of an accurate frame, e.g., about the development of a field, or the personal career path of an influential author. The classifier creation approach of Chapter 5 actively helps filtering data instances. It visually separates instances that support the current hypothesis about the data, and others that do not support this hypothesis. It thus actively aids users discover data instances that lead them to either question or elaborate frames. Both these forms of sensemaking, according to Figure 2.4, lead to a re-evaluation and, if necessary, an update of the current frame. In addition, by storing and visualizing the analysis history as a tree, the approach actively supports re-framing operations by allowing to review past decisions and corresponding, relevant data points. It also facilitates the tracking and comparing of multiple frames by allowing users to store these parallel paths of analysis and seamlessly jump from one to the other. In summary, these approaches mainly support human sensemaking by giving as much control to users as possible, and thus freely allow to steer exploration and analysis to align with their cognitive processes of sensemaking.

In addition, there are aspects of sensemaking that seem particularly critical to visual data analysis. These are, firstly, fixation errors [Klein et al., 2007] that are caused by the expectations about data that active frames generate. Users are then prone to fixate on data consistent with the frame, rather than on data that is inconsistent with it. Secondly, this may go as far as to actively distorting data to match a frame, which seems to be a helpful human trait to be able to handle real world situations effectively and accurately. Thus, this effect may be appropriate and might even be exploited for VA approaches to real time data analysis and situational awareness [Thom, 2015]. For other scenarios, however, in which time pressure is less of an issue, this is surely not conducive to a thorough analysis. How to effectively counteract this phenomenon is unclear currently and future scientific inquiry is needed to produce more insight into this. It might indicate, however, that approaches should not only support correlation of similar data aspects, but more actively facilitate the comparison and analysis of conflicting or contradictory fragments in the data.

Finally, there is still the open question of how humans handle data and frames in order to fit them together during analysis. According to Klein et al. [2007], sensemaking relies heavily on the mode of inquiry of abduction and to a lesser extent on deduction. These two modes of operation, together with a third one called induction, have been discussed by Peirce [1878]. All of them play a role in visual interactive data analysis according to Pike et al. [2009], who identify the relationship between these three modes of inquiry and interaction methods as a major research challenge for VA. Deduction [Johnson-Laird, 1999] is the drawing of syllogisms, and thus logically correct conclusions from a knowledge base. This occurs during sensemaking, e.g., when data is correlated based on event flows known to the user. Insight derived by deduction is thus always correct, given that the premises are correct. Induction is the inference of knowledge based on statistical evidence. For example, if during the analysis of a scientific literature data set, an author appears many times in conjunction with a certain topic, we can assume that this is the author's main research topic. Abduction, which Peirce [1878] calls hypothesis in his early work [Burks, 1946], does not allow to draw logically correct conclusions either. It is the explanation that humans create for surprising facts whose causal context they do not know. For example, the user sees during analysis that an author at a certain point in time starts to publish much more per year and in a much broader set of topics. She thus hypothesizes that this author has either gotten her own working group, or has significantly enlarged her existing team. Creating such hypothesis creates new insights about the data that can either be confirmed or dismissed later. Thus, as Ho [1994] puts it "abduction creates, deduction explicates, and induction verifies." For data exploration abduction plays an important role, as the hypothesis and insights it generates lead to questions that

further analysis based on the remaining two modes of inquiry can try to answer. It thus provides a starting point for the analysis of unknown or partially unknown data sets.

2.2 Natural Language Processing

Despite being older than the research field of visualization, Natural Language Processing (NLP) is a fairly young discipline which is concerned with the automatic processing of human language. It started as an active area of research in the late 1940s, spurred by the then budding Cold War and the prospect of automating Russian to English translations [Jones, 1994]. Since then, its scope has slowly widened, from question answering systems and semantic representations to information extraction and automatic text generation [Jones, 1994]. NLP can be considered a confluence of several disciplines and traditions, all investigating some aspect of language, stretching from linguistics and cognitive science to computer science and electrical engineering. Having split into proponents of two different paradigms for some time [Jurafsky and Martin, 2009], the symbolic faction and the statistical one, the field reunited in the 1990s. NLP approaches according to the former paradigm are based on elaborate linguistic theories that are encoded into intricate rule-based language processing system. The latter paradigm, in contrast, uses flat linguistic representations and large text collections from which statistical machine learning methods automatically induce patterns of language phenomena and use. Nowadays, NLP approaches often combine both paradigms, as suggested by Abney [1996]. Depending on the task they aim to solve, they are equipped with linguistic rules and structures that provide them with information relevant for the task. On top of that, machine learning algorithms infer how to solve the tasks based on the linguistic information provided, as our use case from Chapter 6 illustrates. Language processing methods developed by linguists and NLP experts are included into text VA approaches to provide support for users with language data analysis. These methods are mostly included as data models within the pipeline of Figure 2.1, but can also serve as support for data cleaning and foraging at the beginning of the pipeline.

Including NLP methods into text VA approaches has different implications compared to fully autonomous processing of text. Here, approaches that employ more shallow linguistic representations are often preferred over those that offer deeper analysis of language structure. An example for this is the continuing popularity of text clustering and topic modeling techniques used in VA approaches [Choo et al., 2013; Liu et al., 2012; Cui et al., 2014; Dou et al., 2013]. This is also reflected in the discussion of our taxonomy in Chapter 8. There are several reasons for this. Firstly, shallow statistical models offer advantages in terms of processing speed,

which is important in the context of large data sets for which many VA approaches are designed. Natural language parsing, e.g., is a costly process and thus mostly unsuitable for scalable analysis of text data [van Ham et al., 2009]. Although preprocessing is an option in such cases, it leads to further challenges with respect to interactivity and waiting times before exploration and analysis sessions can be held. Secondly, statistical models are more suitable for user integration, as they typically offer uncertainty information with their results. This can provide useful feedback about the reliability of results, and, if supported by the visualization, information to steer the processing model. Coupling statistical models that operate with shallow linguistic descriptions, such as token lists generated from raw texts enable non-linguists to understand the outcome of the methods. Here, examples from the data and statistics about them can be used to explain automatic decisions to users through the visualization. This also enables user feedback and, very important in VA contexts, steering. Thirdly, VA approaches are designed to have humans as an integral part of their analysis loop. For this reason complex information does not necessarily have to be decode and analyzed automatically. This task can be fulfilled by users through direct access to the text sources, which should be provided by all VA systems. For example, while automatic algorithms can quickly and accurately identify mentions of the same two companies within a huge corpus, human users can use these results to analyze the relevant text passages to learn about the nature of the link between both companies. For these reasons, the methods employed within the context of this thesis are strongly based on statistical techniques. We subsequently discuss them.

2.2.1 The Two Level Model

We see the integration of deeper linguistic representations within VA language processing approaches, to support human users with more tasks than is possible today, as a future challenge for VA and NLP research. For example, the user from the example mentioned above could be supported by a list of automatically generated suggestions about the links between both companies, together with corresponding references to the source text. Chapter 6 discusses an approach that offers support for NLP experts to create accurate linguistic representations for subsequent automatic classification. It lets them explore the state of their machine learning model and helps them discover problematic instances for closer inspection. Visual interactive analysis methods to explore linguistic data is another way to support linguists with their work. The recent years have seen an upsurge of such specialized approaches, e.g., Zhao et al. [2012] and Schätzle and Sacha [2016]. In both cases, linguists and NLP experts are supported in exploring and analyzing their data, and derive knowledge from it. This, in turn, can then be reapplied, either indirectly, through the linguistic phenomena uncovered or theories created.

Or it can be applied directly, through NLP algorithms and toolboxes developed by them.

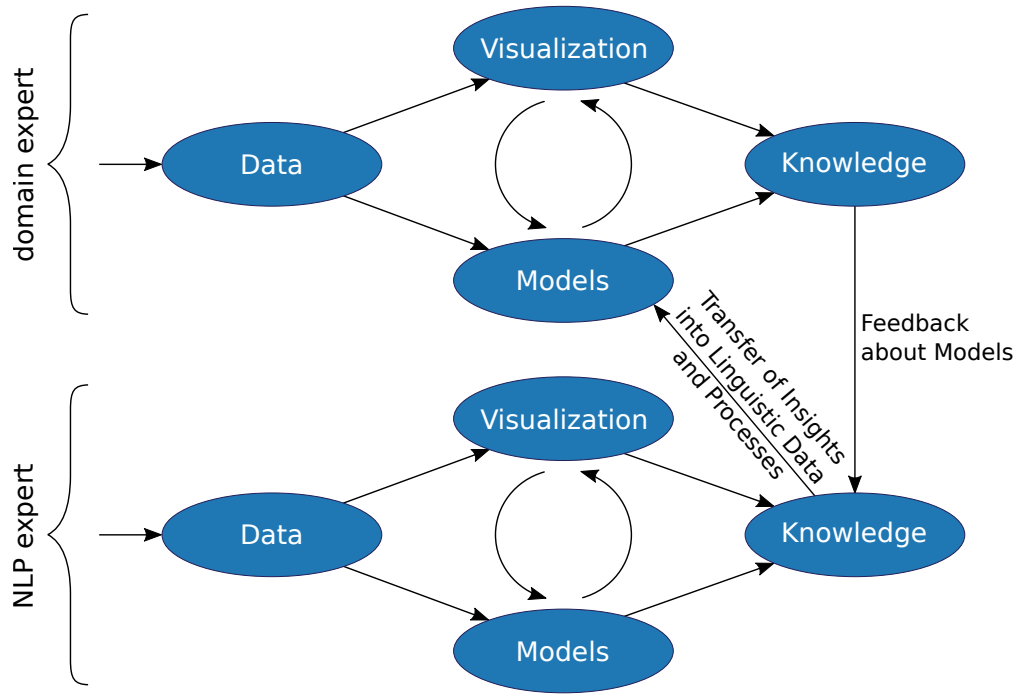


Figure 2.5 — The two level model of language data VA. On the lower level, VA approaches help linguists analyze language data and use their gained knowledge to devise specialized processing models that can then be embedded within VA approaches. These, in turn, help domain experts analyze their text data sets.

The interaction between those two analysis loops are depicted in the two level model for language data analysis (Figure 2.5). Both loops are based on the VA model by Keim et al. [2008]. The lower level represents the knowledge generation process of the NLP experts or linguists during language data analysis. Support for this level is still scarce, but new approaches are being developed continuously, including the one discussed in this thesis (Chapter 6). The upper level describes the text analysis loop of domain experts that explore and analyze text data sets of interest to them. There are two links between both levels. The one that links the knowledge created at the lower level with the models of the upper level signifies the transfer of the knowledge and insights gained by the NLP expert. This can either happen directly, in case the insights from the exploration and analysis process are included into NLP methods or resources, such as dictionaries. Alternatively, this transfer can happen indirectly through insights that are published or otherwise distributed and used by third persons to create analysis methods. The second link, from the domain

expert to the linguist, signifies the transfer of knowledge about the exploration and analysis goals. This can be part of a requirement analysis by NLP engineers, but it can also be informal feedback about problems or benefits of a certain method or system. Of course, the knowledge about the structure of language thus generated is also useful for other purposes besides creating language processing resources for text VA approaches. However, new and more powerful language processing methods, that help support the knowledge generation and sensemaking processes of users, would be an advancement to the current state-of-the-art in text VA. Those would have to satisfy the requirements discussed above, which are differing from the ones for automatic processing, especially with respect to scalability and interpretability of their output. Close collaboration between VA and NLP researchers can thus help to advance both disciplines.

2.2.2 Machine Learning

Machine learning (ML) plays an important role in NLP, because it is able to automatically abstract patterns of language use and structure from a large corpus of sentences or documents. Such algorithms can be fed with linguistic information of various depth about the examples to learn from, depending on what is necessary to accurately solve a learning problem. Generally, ML can be described as function estimation that derives a function $F(\vec{x})$ from n training examples $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$. Each example is represented as a vector of values that describe them in a way that contains sufficient information for the problem at hand. These values are typically called *features*. An ML algorithm, during training, tries to identify the parameters of a previously defined target function $F'(\vec{x})$ in a way that best explains the training examples [Mitchell, 1997]. $F'(\vec{x})$ can, for example, be a simple linear function:

$$F'(\vec{x}) = p_0 + \vec{p} \cdot \vec{x}$$

Then the optimal parameters p_0 and $\vec{p}$ for a given training data set have to be identified, resulting in the function $F(\vec{x})$ that can be applied to any new example. The set of parameters identified for a given training set is called *hypothesis*, and the process of finding a good hypothesis that models the training data well is called *hypothesis selection* [Mitchell, 1997], or simply *training*.

Text Classification and the Vector-Space Model

One classic learning problem in NLP that uses a straightforward feature representation for the training examples is text classification. Here, an ML program learns how to classify documents into different, predefined groups. For example, the training examples could be a set of newspaper articles that are labeled according to whether they are sports news, or any other type of news. Such a classification

problem is called *binary*, because it has exactly two labels for the instances, *sports*, and $\neg$*sports*. In contrast, classification problems with more then two classes, are called *multi-class* problems.

A popular model for creating feature vectors for text classification is called the *vector-space model*. It stipulates a high-dimensional vocabulary space, in which each of the terms that occur within a corpus define one dimension. The typical magnitude of the dimensionality of such vector spaces, depending, among other factors, on the homogeneity of the documents, is in the order of $100,000$. Within this vector space, documents can be represented as vectors according to their term distributions. These document vectors are typically very sparse, as each document only contains a fraction of the possible vocabulary. An advantage of this representation, which makes it suitable for the application of ML algorithms, is that it allows to handle documents with vector algebra. For example, pairwise document distances can be computed based on the Euclidean distance or the cosine similarity between document vectors.

Different weighting schemes have been used over the years to map documents into this space. We briefly mention the three most common:

- **Boolean**
  The Boolean weighting schemes produces Boolean vectors that indicate the presence or absence of terms within a document, ignoring occurrence frequencies.

- **term frequency**
  The term frequency (tf) weighting scheme produces vectors of integers. Each value indicates the frequency of occurrence of the respective term in the corresponding document.

- **term frequency–inverse document frequency**
  The Boolean and the tf schemes both have a significant drawback. They consider every term equally relevant, no matter how typical or frequent it is for the domain of the document set. This introduces a high minimum similarity value between two documents in a corpus due to high frequency terms. The relevant similarity range, in which documents differ, is thus reduced, causing effects such as noise to have more damaging influence. To reduce this effect, tf-idf [Salton and Buckley, 1988] multiplies tf with a correction factor (idf) based on a measure of term popularity. For this, the document frequency (df) of a term, which is the number of documents from document set D it occurs in, is used. Based on this, the inverse document frequency (idf) factor of a term is defined as $idf_{term} = \log \frac{|D|}{df_{term}}$.

A number of additional term weighting schemes have been proposed [Manning et al., 2008; Chuang et al., 2012a]. As we did not use any of those within the context of this thesis, we will not discuss them here. The vector-space model does not preserve word sequences from the original documents. Such schemes are called *bag-of-word* models, since they treat documents as unordered collections of terms. Despite the significant loss of information that is introduced with using such models, they have proven to be effective in many practical applications, such as text classification and information retrieval [Manning et al., 2008].

Different Types of Machine Learning

According to Abney [2007], learning problems can be grouped into four different categories, depending on the co-domain of the function $F(x)$, and the additional information given about each of the training instances. $F(x)$ can either be a nominal function or real-valued [Abney, 2007]. The training instances are assigned values $l_1, l_2, \ldots, l_n$, which are either class labels in the case of a nominal function, or real-valued numbers. Alternatively, each instance is just represented by its plain feature vector without additional information. In the first case, the training of the learning algorithm is called *supervised*, and *unsupervised* in the latter case [Abney, 2007]. If the estimation function is nominal, and a predefined number of classes exist, and each of the training instances is assigned to at least one of them, then the learning problem is called *classification* [Abney, 2007]. An example for this is the text classification example from above. We have used document classification methods within this thesis, and we discuss it in Chapter 5. If the estimation function is nominal, and there are no predefined classes, the learning problem is called *clustering* [Abney, 2007]. Such methods group data samples into meaningful groups based on their feature representations. We use clustering to support several analysis task within this thesis, particularly in the approach discussed in Chapter 7. If the estimation function is real-valued, and the learning algorithm is supervised, the problem is called *regression* [Abney, 2007]. Mühlbacher and Piringer [2013] have presented a VA approach for interactively creating and evaluating regression models. Lastly, if the estimation function is real-valued, and the learning algorithm unsupervised, the learning problem is *density estimation* [Abney, 2007]. It is used, for example, by Thom et al. [2012a] in their VA approach for micro blog message analysis to compute geographic usage densities for terms, allowing them to identify unusual term appearances in geographic space.

Evaluation Criteria

The choice of evaluation scheme for the results of ML algorithms depends on the type of algorithm, the application area, and the data that is available for

evaluation. In this thesis, we primarily focus on classification methods, and for this reason concentrate on their evaluation. For other ML types, other methods are suitable, including Rand [1971] and Färber et al. [2010] for clustering, Montgomery et al. [2012] for regression analysis, and Rajagopalan and Tarboton [1997] and Hall et al. [2011] for density estimation. In addition, the NLP community has developed special evaluation measures that capture the requirements of specific NLP tasks better than more generic ones. Two examples for such tasks and their specialized evaluation measures are machine translation [Han and Wong, 2016], and co-reference resolution [Cai and Strube, 2010].

When applying binary classification algorithms to a new set of instances labeled with ground truth data, there are four different types of possible results for each instance. An instance can be correctly classified as being part of a class, it is then a *true positive* (tp). It can also be correctly classified as not being part of a class, then it is a *true negative* (tn). Finally, it can be incorrectly classified as either being part or not being part of a class, which make it a *false positive* (fp), or a *false negative* (fn), respectively. One measure of classifier performance that is often used is *error rate*. It is defined as

$$error = \frac{fp + fn}{tp + fp + tn + fn}$$

and gives the relative rate of misclassified instances. The opposite of the error rate, called *accuracy* is also a popular measure for classification performance. It is the relative rate of correctly classified instances, and is defined as

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}.$$

Another popular measure in NLP, which we also use in this thesis to evaluate one important aspect of our approach presented in Chapter 5 is the *F measure*. It has been developed for information retrieval [Manning et al., 2008] and is capable of balancing skews in the size of classes, by which accuracy and error rates are negatively effected [Manning and Schütze, 1999]. *F measure* is defined based on *precision*

$$precision = \frac{tp}{tp + fp}$$

which is the relative number of positively and correctly classified instances, and *recall*

$$recall = \frac{tp}{tp + fn}$$

which is the relative number of positive instances that a classifier is able to identify. The final classification quality score is the combination of precision and recall

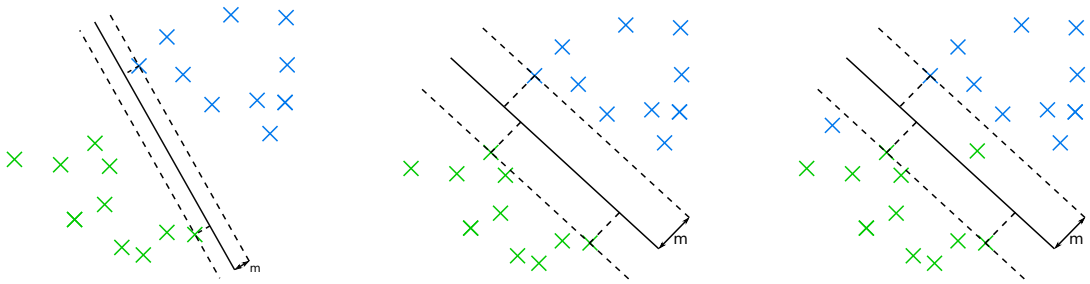
$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

This measure, called F_1 weights recall and precision evenly. Although for information retrieval applications it may be useful to shift the balance between precision and recall, in this thesis we only use F_1. The F_1 measure is also defined for multi-class classification scenarios, such as the use case from Chapter 6. In such cases, two variants are available. All tp, tn, fp, and fn counts can be averaged for all classes, or, alternatively, the precision and recall values are averaged. The first variant, *micro-averaging*, emphasizes the single instances, whereas the second variant, *macro-averaging*, emphasizes the classes [Manning and Schütze, 1999]. In Chapter 6 we use both variants.

A typical strategy to evaluate the classification capabilities of specific algorithms and parameter sets with these measures is cross validation. For cross validation, the set of training instances is split into a larger data set used for training, and a smaller set used to evaluate the resulting classifier. Often, this process is repeated several times, which is then called *k-fold* cross validation [Mitchell, 1997]. Here, the training set is split into k subsets. The training and evaluation process is then repeated k times, once on each combination of $k-1$ subsets as the training set, and the remaining single subset as the evaluation set. Results can then be compared and evaluated statistically. This strategy is also known as *leave-one-out* [Manning and Schütze, 1999]. We adapt it in this thesis to evaluate our interactive classifier creation approach in Chapter 5.

Support Vector Machine

Support Vector Machines (SVMs) [Burges, 1998] are a classification method. They have been introduced by Vapnik [1998], and they are able to learn to distinguish between two classes of data items. Based on identifying a good boundary for two classes of data points, in their basic form, SVMs identify a linear boundary between both classes in the instance space. The resulting classifier is a hyperplane whose dimensionality equals the one of the feature representation of the example instances. SVM uses an optimality criterion for these hyperplanes. It selects the one, for a given training set, that maximizes the space between the separating hyperplane and the closest training instances on both sides. Figure 2.6 depicts a simple example in 2D space. The learning algorithm's task is to separate the green from the blue entities. While the first configuration (Figure 2.6a) shows an arbitrary separating hyperplane for both classes, the second one (Figure 2.6b) depicts the plane with the largest possible space between the separator and the training examples next to it. The distance from the separator to the nearest data points, m, is called the margin of the hypothesis. An SVM's goal is thus to maximize the margin for the separator of a given data set. This is tantamount to selecting the least risky hypothesis, keeping the "safe space" as large as possible [Abney, 2007].



(a) arbitrary margin (b) largest margin (c) non-separable case

Figure 2.6 — Two possibilities of linear separation the green training examples from the blue ones, loosely base on the examples by Abney [2007]. One of them separates the data set with an arbitrary margin m (a), while the other example uses the largest possible margin (b), which is the optimality criterion of an SVM. The third example shows the maximal margin for a non-separable data set, requiring the use of slack variables.

Once induced from a training data set, an SVM classifier can then be applied to new instances and estimate their class membership based on which side of the hyperplane they lie. A hyperplane can be defined by its distance from the origin, p_0, and its orientation in the feature space, expressed by its normal $\vec{p}$. The classification function $F(\vec{x})$ for an SVM is thus

$$F(\vec{x}) = \operatorname{sgn}(\vec{p} \cdot \vec{x} + p_0)$$

which evaluates to +1 or −1 depending on the side of the hyperplane example $\vec{x}$ lies on. Defining the class labels of the training instances $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ as $l_1, l_2, \ldots, l_n$ with $l_i \in \{-1, +1\}$, the distance d_i of each instance $\vec{x}_i$ to the hyperplane is given by

$$d_i = l_i \cdot \frac{\vec{x}_i \cdot \vec{p} + p_0}{|\vec{p}|}$$

The margin m is thus defined as the minimal distance from a training data point to the hyperplane: $m = \min_i d_i$. Formulating the optimization problem, and applying Lagrangian multipliers (α_i) to it leads to the following dual problem

$$\begin{aligned} \underset{\alpha_1,\ldots,\alpha_n}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} l_i l_j \alpha_i \alpha_j \vec{x}_i \vec{x}_j \\ \text{subject to} \quad & \sum_{i=1}^{n} l_i \alpha_i = 0; \alpha_i \geq 0, i = 1, \ldots, n. \end{aligned}$$

with the parameter vector being a linear combination of the training examples $\vec{p} = \sum_{i=1}^{n} l_i \alpha_i \vec{p}_i$, the margin $m = 1/(\vec{p}^2)$, and $p_0 = -\frac{1}{2}(\max_{l_i=-1}(\vec{p}\vec{x}_i) + \min_{l_i=+1}(\vec{p}\vec{x}_i))$ [Cristianini and Shawe-Taylor, 2010]. Depending on the training data set, for many

training examples $\alpha_i = 0$, which means that they have no influence on the hyperplane. Those examples, for which $\alpha_i > 0$, are called the *support vectors*. They lie closest to the boundary, and defined the separating hyperplane.

Of course, SVMs are also applicable to training data sets that are not linearly separable, such as the one depicted in Figure 2.6c. In this case, the weights for the examples have to be constrained to a constant C, such that $C \geq \alpha_i \geq 0$ for $i = 1, \ldots, n$ [Cristianini and Shawe-Taylor, 2010]. This limits the possible influence of single examples on the hyperplane, and thus keep the optimization problem solvable. A good choice of C is dependent on the data set and its feature representation, and can be experimentally determined. Although being a linear classifier in its basic form, one property of SVMs is that they can be easily extended to a vast range of different classification functions through the so-called *kernel trick* [Manning et al., 2008]. As the instance vectors within the SVM framework only appear as scalar products, these can be replaced by a kernel function. Such a function, that has to satisfy certain conditions [Manning et al., 2008], represents an implicit mapping into a higher-dimensional feature space and the computation of the scalar product in this feature space. Although linear SVMs are known to work well for text classification [Joachims, 1998, 1999], we use this trick in Chapter 5 to accurately classify very short micro blog messages that exhibit a high level of noise with the help of a specialized kernel.

Active Learning

One of the downsides of classification is the need for labeled training data. For language data, this typically involves significant human effort to prepare labeled instances. To overcome this hurdle and reduce the need for human labor when creating high-quality classifiers, a number of techniques have been proposed. There are basically two types of methods. One is called *semi-supervised learning* [Abney, 2007] and is based on the idea of using a small set of labeled instances, and a larger set of unlabeled ones. Such algorithms are capable of using the small number of labels efficiently by extracting additional information from the large number of unlabeled examples. The second type of approaches is called *active learning* (AL) [Settles, 2012]. They also start with a small initially labeled set of instances. Then, classifiers are created over multiple iterations, identifying one or more unlabeled examples during each one that are expected to contain the most information for the training algorithm. These instances are then presented to human annotators that label them according to the classification task. They are subsequently added to the training examples and a new AL iteration starts. Thus, labeling effort is reduced by selecting the most informative examples to label during each iteration [Settles, 2009]. Figure 2.7 depicts the AL loop according to Settles [2009].

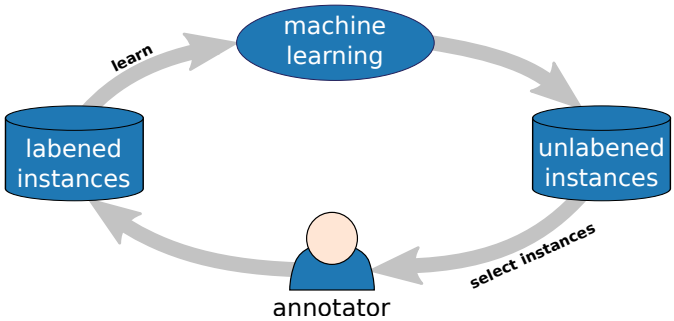


Figure 2.7 — The active learning loop, adapted with modifications from Settles [2009].

Data: labeled training set D_l and unlabeled data D_u
Result: labeled data D_l and classifier C that is trained on D_l

1 apply base learner L_b to D_l to obtain classifier C;
2 **while** *stopping criterion is not met* **do**
3 apply C to D_u to obtain D'_u;
4 select the most informative n instances D_i from D'_u;
5 ask annotator for labels of instances in D_i;
6 move D_i with labels to D_l;
7 apply B to D_i;
8 **end**
9 output D_l and classifier C;

Algorithm 1: The basic active learning algorithm based on Olsson [2009].

Olsson [2009] describes this process in the form of a prototypical algorithm (see Algorithm 1). The interesting part of Algorithm 1 is how informativeness in line 4 is measured. There are multiple schemes that have been proposed for this. The one that we base our visual and interactive approach from Chapter 5 on is called uncertainty sampling [Lewis and Gale, 1994]. It is based on the intuition that classifiers can learn more from examples they are uncertain about, rather than those that they process with high confidence. Uncertainty sampling has been reported to significantly reduce labeling effort for text classification [Lewis and Gale, 1994]. This is also true for SVMs [Schohn and Cohn, 2000; Campbell et al., 2000; Tong and Koller, 2001], which we use for our approach in Section 5. Here, sampling the instances with highest classification uncertainty is tantamount to selecting those closest to the hyperplane. Tong and Koller [2001] show that this method reduces the space of classification hypotheses as fast as possible, converging to the optimal hyperplane with significantly reduced labeling effort. We exploit this principle to visually guide users to the most advantageous instances to label. In contrast to

this classic algorithm, our approach entirely hands off control over which instances are selected for labeling to the user.

Spectral Clustering

The next method we discuss is a clustering approach for textual and other data instances. While designing our methods from Chapter 7, we had the requirement to support the interactive exploration of a scientific literature set over time. Based on the decision about the visual method to convey the contents to users (see Chapter 7 for details), and allow them to be inspected in an interactive way, we decided to use a clustering method to automatically organize the data set. Another requirement of the approach from Chapter 7 was that users should be supported with abstracting from local properties and explore higher-level structures, while also having access to local dynamics of the data. Based on these requirements, we decided to use the spectral clustering technique [Von Luxburg, 2007]. It is a top-down hierarchical method which generates a cluster tree instead of a flat set of clusters. This lets users interactively adapt the number of clusters by splitting them without modifying their boundaries, which is conducive to creating and keeping a mental map of the data set. Thus, users can adapt the granularity of the view on the text data. In addition, the cluster tree is binary, i.e., exactly one cluster is split at each level of the tree, allowing for a very fine-grained adjustment of the cluster granularity. A further advantage is that spectral clustering is not biased with respect to the shape of the clusters that are generated and is thus capable of yielding more natural clusters compared to other methods.

The spectral clustering algorithm recursively searches for the best cuts in a neighborhood graph, i.e., a set of severed edges in a graph that partition the set of nodes. For this, the instance set is transformed into a k nearest neighbor graph, for which we use $k = 10$ for our documents. We determined this value using Von Luxburg [2007]'s method. One criterion used to identify the best cut is high intra-cluster and low inter-cluster similarity. This is achieved by severing edges whose sum of weights is as low as possible. Using this objective alone would split off many small clusters that have weak links to the remaining instances. To solve this problem, a second optimality criterion that balances cluster size is typically included. An objective function that incorporates both criteria is the normalized Cheeger cut [Cheeger, 1970]. We have decided to use it, as recent results suggest that it outperforms other objectives [Bühler and Hein, 2009]. The normalized Cheeger cut (NCC) is defined as in Equation 2.1. C and $\overline{C}$ represent two candidate partitions of the graph during an iteration of the algorithm, and w_{ij} is the similarity between node

i and j. Equation 2.2 defines the cut size, and Equation 2.3 is a measure for the size of a graph's partitions.

$$NCC(C,\overline{C}) = \frac{cut(C,\overline{C})}{min(vol(C), vol(\overline{C}))} \tag{2.1}$$

$$cut(C,\overline{C}) = \sum_{i \in C, j \in \overline{C}} w_{ij} \tag{2.2}$$

$$vol(C) = \frac{1}{2} \sum_{i,j \in C} w_{ij} \tag{2.3}$$

Finding a solution that minimizes Equation 2.1 is an NP-hard problem. Luckily, it can be efficiently approximated using spectral methods. Bühler and Hein [2009] show that an approximate solution for the optimal Cheeger cut can be found through the second smallest eigenvalue of a Laplacian matrix of the neighborhood graph. We use their implementation of this technique to apply this clustering methods to our data.

2.2.3 G^2 Keyword Extraction

Further statistical techniques in addition to ML exist in NLP that are useful for creating text visualization approaches. In some of our approaches, namely those from Chapters 4 and 7, we had the problem of identifying keywords that accurately characterize a subset of documents from a larger set. While the tf-idf measure is one possible option for this problem [Viégas et al., 2006], Chuang et al. [2012a] show that it is significantly outperformed by the G^2 measure in yielding terms expected by humans. This is due to the fact that the latter is based on the discrepancy of occurrence frequencies between the subset and the entire document set [Rayson and Garside, 2000]. G^2 therefore has the added benefit of yielding statistical significance values for the difference between a term's occurrence frequencies on both sets [Collins et al., 2009c]. The measure is calculated by counting the occurrences of each term w in each of the two document sets, which results in a contingency table as shown in Table 2.1. Based on the frequency of a term in the entire data set, expected frequency values for each of the two subsets are computed (Equation 2.4).

$$E_1 = c \cdot \frac{a+b}{c+d} \quad \text{and} \quad E_2 = d \cdot \frac{a+b}{c+d} \tag{2.4}$$

$$G^2 = 2a \cdot \log \frac{a}{E_1} + 2b \cdot \log \frac{b}{E_2} \tag{2.5}$$

The difference between expected and real frequency values for term w is tested for

	subset	¬subset	total
freq t	a	b	a+b
freq $\neg t$	c-a	d-b	a+d-a-b
total	c	d	c+d

Table 2.1 — Contingency table for term t, adapted from Rayson and Garside [2000]. The columns *subset* and *¬subset* denote the relevant subset and the remaining set, respectively.

statistical significance under the null hypothesis that differences are of a purely random nature (Equation 2.5). The G^2 score thereby approximates a χ^2 distribution that makes it possible to derive probability values for the null hypothesis using the standard χ^2 calculators or tables. Higher values of G^2 correspond to lower probabilities for the null hypothesis, allowing us to rank and size terms, e.g., in word clouds according to that value.

2.3 Visual Analysis of Text and Scientific Literature

This chapter contains previously published material from the following publications:

P. Federico, F. Heimerl, S. Koch, and S. Miksch. A survey on visual approaches for analyzing scientific literature and patents. *Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. doi: 10.1109/TVCG.2016.2610422

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Transactions on Visualization and Computer Graphics*, 18 (12):2839–2848, 2012b. doi: 10.1109/TVCG.2012.277

F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word Cloud Explorer: Text analytics based on word clouds. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1833–1842, Jan 2014. doi: 10.1109/HICSS.2014.231

In this section, we give an overview of text analysis and exploration approaches that have been applied to scientific literature data. For this, we first take a brief look at the state-of-the-art in text visualization in general. Then, we focus particularly

on those approaches that are designed for scientific literature, and discuss them in more detail. For all other visualization approaches to scientific literature that do not include any text analysis aspect, we refer the interested reader to Federico et al. [2016]. This section gives a broad overview of the field that this thesis is rooted in. For a discussion of related work specific to the single approaches of this theses, each of the following chapters comprises a corresponding separate section dedicated to this.

2.3.1 Overview of Text Visual Analytics

Although still very young, VA research of text exploration and analysis approaches has gained an impressive momentum during the past decade. Early approaches have often concentrated on 2D document spatializations to convey content similarities through spatial proximity to users [Wise et al., 1995; Chalmers and Chitson, 1992; Olsen et al., 1993]. Through such approaches, which we employ, e.g., in Chapter 4, users can explore the structure of a document set based on visual clusters. Since then, many approaches that support different sets of analysis and exploration tasks have been proposed. ThemeRivers [Havre et al., 2002] have been devised to depict temporal dynamics of word occurrences. They help understand trends and theme developments in a corpus over time. *Jigsaw* [Stasko et al., 2008] offers a variety of integrated visual tools to relate entities from documents. Parallel tag clouds [Collins et al., 2009c] makes use of multiple word clouds to explore and analyze term use in corpora with different facets. Oelke et al. [2010] provide an interactive approach that lets authors analyze their writings with respect to readability. To integrate users more deeply into the analysis process, the concept of semantic interaction has been introduced. It allows domain experts to interact with data models through the visualization of the data rather than having to interact with it directly [Endert et al., 2011, 2012]. More recent text VA approaches focus on abstract visual representations of large document sets to help understand the types of information they contain. For this, topic modeling, most prominently Latent Dirichlet Allocation (LDA) [Blei et al., 2003], has become a popular method to aggregate large document sets based on their term distributions. One of the first visualization approach to use topic modeling is Liu et al. [2012], who base their visualization on ThemeRivers to convey temporal topic dynamics. This visualization paradigm has become a popular choice for temporal topic dynamics and has been adapted in many of the subsequent approaches. Two recent approaches by Dou et al. [2013] and Cui et al. [2014] allow users to visually explore a hierarchy of topics derived from a text data set, offering drill-down interaction. Other approaches focus on text collections with heterogeneous documents of different classes or sources whose correlation with the depicted topics can be analyzed [Oelke et al., 2014; Liu et al., 2014]. Some authors have focused on VA of social media data. They have

created techniques that let users explore the interaction of topics, or the diffusion of information in online platforms [Xu et al., 2013; Sun et al., 2014; Wu et al., 2014; Zhao et al., 2014]. Alexander and Gleicher [2016] propose an approach that facilitates the comparison of multiple topic models with respect to their suitability for different analysis tasks. Recently, topic modeling techniques have been applied to interactive visual debate analysis [El-Assady et al., 2016]. Over the past recent years, new and exciting approaches beyond topic modeling have also been appeared for exploration and analysis tasks of scholars from the humanities, e.g., by Koch et al. [2014].

2.3.2 Text Visualization for Scientific Literature

Organizing, analyzing, and exploring human knowledge has been an active research area for a long time. Researchers from many fields have devised methods to categorize and make sense of the massive amounts of available scientific writings. Information science is a field particularly devoted to developing data analysis methods for this goal. Examples of such methods include measures of the prolificacy of authors [Hirsch, 2005], or the identification of research fronts [Garfield, 1994], i.e., current cutting edge research topics. In addition, the data mining community has also developed techniques for scientific literature analysis, from new search methods [El-Arini and Guestrin, 2011] to summarization approaches for entire disciplines [Shahaf et al., 2012]. Also, many techniques exist for the visualization of bibliographic data, which often produce visually appealing, static images of multiple disciplines of science called science landscapes [Börner, 2010]. In this section, we focus on text visualization approaches that are either designed for or applied to scientific literature.

The first category of approaches that we look into are designed to visually support retrieval of scientific documents. *TileBars* [Hearst, 1995] is a visualization for search results that is based on a text tiling technique to split texts into coherent thematic sections. It depicts a visual summary of the distribution of search terms across the sections of a document as a rectangular strip. Through the length of the strip users can compare relative document lengths and differing term distributions in the result set. Nowell et al. [1996] present a very flexible way of organizing search results. Users can organize and explore result sets in the cells of a matrix view whose axes can be set to any aspect of the documents, e.g., author and year of publication. Additional metadata attributes can be included by mapping them to shape, pictograms, label, or color of the respective document icon in its matrix cell. Koch et al. [2011b] introduce a VA approach for query extension and refinement of Boolean queries. It supports the exploration of result sets through multiple linked views of the distribution of different metadata aspects. Other approaches allow

users to define multiple queries and visualize documents based on their relevance to them. Olsen et al. [1993] propose the *VIBE* system that positions documents on a 2D plane relative to multiple queries. These queries can be defined and positioned freely by users, which results in a highly interactive system for exploring document sets. Scalability is limited, however, as the positions of the documents become ambiguous for four or more queries. *GUIDO* [Korfhage, 1991] is a similar method that does not support free query placement. It optimizes query position and maps documents according to their absolute distances to queries. Although this theoretically introduces less information loss, the resulting complex geometric forms are harder to interpret. Sparkler [Havre et al., 2001] uses a different spatialization scheme to facilitate the comparison of multiple queries. It distributes result sets on a circle split into one segment per query. Documents can occur in multiple sets, and their glyphs are colored according to the query. Distance from the center encodes relevance to a query. To avoid overlaps, documents are spread out radially within the boundaries of the segment. This gives users an overview of the relevance distributions for each query and helps to compare results. Costagliola and Fuccella [2011] also spatialize search results for a query. They are laid out according to their textual similarities in a circular area. This area is further extended into a 3D tube by adding a third dimension for the publication time of the articles. The interface supports the standard 3D interactions to combat occlusion problems. References between the articles in the set can be displayed as edges on demand.

The next category of approaches are designed to explore and analyze relations between documents or other objects. Strobelt et al. [2009] represent each publication with a card that contains descriptive terms and representative images as a brief summary of its content. This enables users to quickly compare documents contents within a collection. Chuang et al. [2012c] present an approach to find topic relations between different university departments based on their doctoral dissertations. For this, they apply topic modeling to them and project the result into 2D using principal component analysis. Such a projection gives an overview of document similarities, but also distorts the original distances. Users can select a single department, which causes the others to be laid out in a circular fashion around it with their true distances to the selected department. This provides users with a way to learn about the distortions of the projection. A method to find and relate thematic clusters in a citation network is presented by Nakazawa et al. [2015]. They use topic modeling to group documents by topic and then depict these clusters as nodes of a graph. Görg et al. [2010] help users find relations between biomedical publications based on the biological entities, e.g., genes, mentioned in their abstracts and index terms. It enables users to correlate and combine findings about interconnections between entities to generate new insights. A more general version of this approach [Görg et al., 2013] allows analyzing publications from

any domain. It combines various text analysis methods, including clustering of documents by topic, and named-entity extraction. Information extracted from texts can be correlated with metadata of the documents. Riehmann et al. [2015]'s approach is designed to relate and compare text from different documents to explore cases of alleged plagiarism. The visualization is based on a bipartite graph, linking text passages in the suspicious documents with possible sources. *Rexplore* [Osborne et al., 2013] is a web-based system for search and faceted browsing of publications that features a node-link graph connecting similar authors. Similarity of authors is computed based on the content similarity of their publications. *PaperLens* [Lee et al., 2005] groups papers by research topics, depicted as bar charts per year. It includes rankings of the ten most highly cited authors per year, and allows users to identify connections between authors in a co-author graph.

The following category of approaches supports the aggregate analysis of patterns based on multiple documents or entire data sets. *VxInsight* [Davidson et al., 1998] visualizes clusterings of documents. The resulting 3D visualizations follow a map metaphor, representing dense areas as hills, and areas of lower density as valleys. Users can zoom and rotate the resulting terrain. The valleys are labeled with representative terms to assist navigation. *IN-SPIRE* [Wong et al., 2004] also enables document set spatialization. It provides two different visualizations, a scatter plot of documents in 2D, and, similar to the previous approach, a 3D metaphor with mountains for dense, and valleys for less dense areas. *GistIcons* [DeCamp et al., 2005] are circular histograms of terms for single documents. To make the resulting, individual shapes comparable, the terms are grouped by concepts. Based on the visual similarity of the shapes, users can identify topics and groups of similar documents. Wu et al. [2011] achieve a similar effect using word clouds based on a new algorithm to optimize them. The *Termite* system [Chuang et al., 2012b] is designed to give insight into automatic topic modeling results. It consists of a term-topic matrix that includes the subset of most distinguishing terms between the topics. The relevance of a term for each topic is depicted by circles of varying sizes. This provides users with an overview of the topics in a document set and their meaning. Chuang et al. [2012d] use this technique for the exploration of topics in a set of PhD dissertations. They combine it with a citation graph along a time line that visualizes the influence from a cited to a citing paper as topic flows. The size of the documents, depicted as nodes of the graph, encodes their overall influence on others. This shows historic developments over time and helps to identify highly influential papers. *ParallelTopics* [Dou et al., 2011] is another approach designed to analyze and explore topic modeling results. It includes multiple views, such as word clouds for each topic, and a streamgraph to depict temporal dynamics. In addition, a scatter plot provides information about the number of topics each document contains, and a parallel coordinates view gives a detailed account of

the topic distribution for selected documents. Jiang and Zhang [2016] also use a topic scatter plot to depict topic similarity. They combine it with a Sankey diagram that shows topic evolution in a data set over time. *Serendip* provides three views that give access to different levels of a corpus, the word level, the document level, and the corpus level [Alexander et al., 2014]. All three levels can be explored in an integrated fashion based on topic modeling results. This provides users with insights into the topic structure of a scientific literature set during exploration. Gretarsson et al. [2012] present a web-based topic exploration approach that shows topic modeling results as node-link diagrams and lets users explore their connection to publications and associated university departments. The *iVisClustering* technique [Lee et al., 2012] uses topic modeling that can be steered by users to visually classify documents based on its results. Oelke et al. [2014] focus on document sets of up to three classes. These classes are depicted by splitting a rectangular area into up to three subareas. Circular document coins that contain word clouds of topics extracted from the underlying documents are placed within or at the border of these areas, and show the affinity of a topic to one or multiple classes. Choo et al. [2013] propose *UTOPIAN*, a system that bases a 2D mapping of documents on topic extraction results. Through this 2D scatter plot, on which clusters are labeled with keywords, users can interact with the topic modeling algorithm. The results can thus be adjusted to the goals of the user.

Maps or landscapes are a popular visualization metaphor for scientific disciplines. Many of the mapping approaches are based on either citation or co-author data, but there are examples of maps based on textual data. Fried and Kobourov [2014] create a map of computer science articles based on titles from a large publication database. Their algorithm extracts keywords, and links them based on word co-occurrences. Users can activate a heat map that highlights certain areas. Thus, profiles of researchers, research institutions, or conferences can be depicted relative to the entire map. Skupin [2002] creates a map based on contents of conference abstracts. The resulting groups are clustered a second time into a hierarchy from which maps of various granularity can be created. Skupin [2004] presents an entire pipeline into which he plugs multiple clustering methods, discussing and evaluating the resulting maps.

The following paragraph is devoted to approaches that focus on the analysis of temporal aspects of scientific document sets. In Mane and Börner [2004]'s approach, temporal dynamics play a central role. They extract important terms from titles and keywords of biological articles and plot their occurrences over time. This allows the detection of bursts caused by sudden interest in a particular theme. Term co-occurrences can also be depicted as a graph, encoding the temporal dimension by the color of the nodes. Creating and visualizing a hierarchy of topics is explored by Dou et al. [2013]. They create topic hierarchies through a combination of topic

modeling and a hierarchical clustering algorithm. The resulting tree of topics is depicted in a node-link fashion and can be modified and adapted according to the user's analysis goal. Topics associated with each node can be explored in a streamgraph view that depicts their development over time. Ahmed et al. [2004] use a 3D approach similar to streamgraphs to visualize clustering results on a data set of publications. Clusters are depicted over time as worm-like 3D structures of varying thickness. Citations between papers in different clusters are depicted by edges between time points of the streams. Liu et al. [2015a] provide an interactive visualization approach that allows users to analyze two text data sets with respect to which of the sets contains innovations that are later taken up by the second one. They base their technique on topic modeling and provide multiple linked views for interactive visual analysis including a topic hierarchy tree and a time line that depicts the propagation of innovation through text corpora. Technically, the method bears some resemblance to our trendiness score described in Chapter 7.

So far we have discussed approaches centered around text retrieval, analysis, and exploration tasks. In addition to those, the community has yielded additional approaches that concentrate on different data types and couple them with text analysis. Some of these approaches use text processing to extract labels for other data entities, such as citation network nodes [Chen, 2006], and citation and co-authorship graphs [van Ham, 2004]. Other approaches feature a deeper integration of textual content and document meta data. Chen [1999], for example, combine latent semantic indexing and co-author analysis to visualize content and co-citations patterns. Their proposed visualization links articles grouped into different categories that correspond to thematic fields. The *ActionScienceExplorer* [Dunne et al., 2012] integrates into the popular JabRef application for reference management and facilitates the exploration of citation graphs. The nodes of the graph, that represent publications, are clustered according to their citation impact. In addition, the contents of multiple publications can be automatically summarized and skimmed by users. Honkela et al. [2011] use SOMs to spatialize scientific documents based on their contents. Publications in different languages can be processed by the approach based on the use of machine translation. Once the document map is created, different metadata aspects, e.g., authors, can be mapped into the document space to view their distribution. Another clustering technique is used by Jusufi et al. [2014]. They group documents by content into clusters depicted as graph nodes, and link them based on co-author information. Thus, research topics shared by authors can be explored. Sharara et al. [2011] present an approach to interactively evaluate graph classification models. Their corresponding implementation is applied to a scientific literature data set on which it is shown to predict topics for papers in a citation graph. *PivotPath* [Dörk et al., 2012] is an interactive data exploration

approach that depicts scientific document collections across three levels that each focus on a different aspect of authors, resources, and concepts. These aspects are linked and their interrelations can be explored based on the highly interactive user interface.

CHAPTER 3

Casual Interactive Analysis of Single Text Bodies

This chapter contains previously published material from the following publications:

F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word Cloud Explorer: Text analytics based on word clouds. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1833–1842, Jan 2014. doi: 10.1109/HICSS.2014.231

S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl. ConcentriCloud: Word cloud visualization for multiple text documents. In *International Conference on Information Visualisation (IV)*, pages 114–120. IEEE, July 2015. doi: 10.1109/iV.2015.30

The first approach presented in this thesis has been designed for the interactive exploration of single, coherent bodies of text. This could be newspaper articles about a certain topic or, in the context of scientific literature, all the publications of a single author. The analysis of one single document, such as a novel, is also possible. This work aims at the design of an interactive interface that casual users [Pousman et al., 2007] can use effectively for text exploration tasks. Its visual

encodings are therefore straightforward to understand, and the interaction concept makes it easy to use without requiring much initial training. This has spurred interest from numerous users in other disciplines, such as researchers from the humanities and biologists, who thought of it as a great tool to explore literary as well as scientific texts. With this approach, we also demonstrate how integrating NLP methods that are accessible through interaction can help casual users in many analysis and exploration scenarios.

3.1 Motivation

Getting an overview of a confined collection of texts quickly is a task that we are often confronted with. "What were the most interesting events during the past Olympic games?" we might ask, or "What has this researcher's work of the past five years been about?". Usually, we would have to skim or read multiple documents of linear text to answer these questions. Depending on the level of detail of the answer that we expect for our questions, we would need a considerable amount of time and effort to explore the collection of documents that contain the answer. Often, topic modeling (see Chapter 2) is used for the purpose of gaining an overview of large document sets. However, to use it effectively, users first have to understand the concept of a topic, and what it represents. This is not possible in a casual exploration scenario. In addition, topic modeling, which is based on a statistical analysis of word co-occurrences in documents, needs more than a few of them to be effective. For this reason, to help with such casual, everyday, exploratory text analysis tasks, we have designed the Word Cloud Explorer. It is an interactive text exploration approach entirely based on word clouds as the main visualization paradigm. Word clouds are a widely known, abstract visual representation of texts that is straightforward to interpret for casual users. They are a visualization technique that has appeared outside of academia [Viégas and Wattenberg, 2008], and they have become popular on the web, in the context of community-oriented websites [Smith, 2008]. Historically, word clouds are sometimes called *tag clouds*, stemming from their early use to depict tag popularity on internet platforms. When used for text visualization and analysis, they are typically called word clouds, and we adopt this terminology here. One prominent application area for word clouds is visual text characterization [Burch et al., 2013; Feinberg, 2010; Kuo et al., 2007]. Here, they are used to give an intuitive and visually appealing overview of a text by depicting the words that occur most often within it. Such a characterization is particularly helpful as a first step for exploratory analysis, because it quickly conveys prominent terms and topics within a body of text. Viégas and Wattenberg [2008] argue that word clouds owe part of their popularity to the fact that they provide "a friendly atmosphere" and "are fun rather than businesslike". This playful

environment they create spurs human sensemaking processes (see Chapter 2), and therefore fits well into a casual text exploration approach.

Word clouds often serve as a starting point for text exploration and subsequent deeper analysis [Burch et al., 2013; Sinclair and Cardew-Hall, 2008; Viegas et al., 2007]. They can, for example, help users to quickly judge whether a given document contains relevant information for them. Typically, this overview is achieved by laying out the high frequency words from a text on a 2D plane, and scale their font size according to their frequency of appearance. One of their drawbacks is that they provide a purely statistical summary of isolated words without taking any knowledge about the words and their relations into account. They thus have limited exploratory capabilities, relegating users to passive viewers of a static display of information. In many approaches that include word clouds, they are used that way and typically provide no or limited interaction capabilities. With this approach, we explore the possibility of using word clouds as the visualization and interaction element at the very center of a text analysis framework. We combine them with interaction techniques that allow to filter and confine the information that users want to explore. The interaction and filtering capabilities are based on NLP methods that correlate, extract, and classify words in the original text body. With the aim of keeping the approach usable by casual users, we decided to include NLP methods according to two aspects. Firstly, their output should be interpretable by anyone without a background in linguistics. Secondly, the accuracy of the approaches should be very high, releasing users of the burden to gauge the correctness of automatic analyses. Word Cloud Explorer includes multiple NLP methods, sophisticated interaction, and a high level of control for users to provide support for a variety of exploration tasks. It is thus not only suited for casual users, but is also highly customizable and thus a powerful approach for experienced users. In the following, we discuss the Word Cloud Explorer and the design choices on which it is based. We then provide a usage example for it and discuss the results of a user feedback session. Finally, we discuss an additional design study to extend the approach to contrastive exploration of small document sets. Due to its circular layout, we call the extended design and its corresponding implementation ConcentriCloud.

3.2 Related Work

Apart from popular platforms [Wordle, 2014; Tagul, 2009; Tagxedo, 2016] that produce customizable but static word clouds for user-uploaded texts, there is considerable scientific work relevant to us. Work on word clouds either investigates the effectiveness of word clouds, improve their creation or layout, or embed them within larger text analysis and exploration approaches. Examples for the latter are

Koch et al. [2011b], who use them to characterize patent sets during retrieval, and Stasko et al. [2008], who use them to support investigative analysis. We also use word clouds in our approach in Chapter 7. In these systems, static word clouds are used to visually characterize texts. An interactive word cloud variant has been implemented by Dörk et al. [2008] for the exploration of web search results. Similar to our approach, the terms in the word cloud can be filtered. However, the approach does not offer the same level of analytical capabilities for exploration than ours, e.g., with respect to interactively filtering words. Vuillemot et al. [2009] include a word cloud for vocabulary analysis in their literary analysis system. It supports filtering by part-of-speech. Alternatively, the characters from a literary text can be displayed in the cloud. However, its analytical and interactive features are limited compared to ours with respect to interactivity and configurability. These make our approach more versatile and applicable to a wider range of text types.

There are several investigations into the effectiveness and the perceptional properties of word clouds. According to Rivadeneira et al. [2007] and Bateman et al. [2008], font size and color has a strong effect on user's attention. In addition, terms in the middle of the cloud receive more attention than terms near the borders [Bateman et al., 2008; Lohmann et al., 2009]. Studies that compare word clouds to unweighted lists of words and other user interfaces [Halvey and Keane, 2007; Lohmann et al., 2009; Rivadeneira et al., 2007] indicate that users are more effective in spotting a specific term in an alphabetically ordered list than in a word cloud ordered according to the same criterion. However, frequently used terms are found more quickly in word clouds due to their larger font sizes [Bateman et al., 2008; Lohmann et al., 2009; Rivadeneira et al., 2007]. Alexander et al. [2016] report that word length has a distorting effect on user's ability to recognize font size differences, but they find the effect to be surprisingly weak. Sinclair and Cardew-Hall [2008] compare word clouds with a user interface that only consists of a search box. While participants preferred the latter to enter specific terms, they favored the word cloud for exploratory tasks. Similarly, Kuo et al. [2007]'s results indicate that word clouds are effective to give an impression of what information is present in a document set, being a good method to convey an "overall picture" of texts. This makes them a good fit for our goal of creating a framework for casual text exploration.

A popular layout scheme for word clouds is the line-by-line layout within a rectangular shape. Several alternatives to this have been studied, with the goal of reducing white space [Kaser and Lemire, 2007; Seifert et al., 2008]. Other approaches optimize word distances to indicate relatedness by their spatial distance [Hassan-Montero and Herrero-Solana, 2006; Schrammel et al., 2009; Chen et al., 2009; Fujimura et al., 2008; Paulovich et al., 2012; Wu et al., 2011]. Word relations can also be encoded by connecting terms with arcs [Stefaner, 2007] or by highlighting related words on demand [Dörk et al., 2008; Lohmann et al., 2012]. Temporal

changes in word usage can be depicted in word clouds by using sparklines [Lee et al., 2010] or histograms [Lohmann et al., 2012]. While these visualizations can be used to illustrate the evolution of words in different text documents, the documents themselves are not distinguished in the word clouds, as we do with the ConcentriCloud approach. The same limitation holds for Cui et al. [2010], who combine trend charts with word clouds to illustrate the temporal evolution of words. Tree Clouds [Gambette and Véronis, 2010] combine word clouds with trees to visualize the semantic relatedness of terms. Prefix Tag Clouds [Burch et al., 2013] use prefix trees to group different word forms and visualize the subtrees as word clouds. We, in contrast, use NLP to automatically merge morphological word variants. While a part of these extensions are designed for specific application contexts, others can be used more generically. We adopted some of these ideas in this approach, such as the circular word cloud layout or the interactive highlighting of term relations.

Some approaches support multiple documents through small multiples of word clouds [Wu et al., 2010; Seifert et al., 2014; Castella and Sutton, 2014]. All of these approaches, however, display the same words multiple times for different documents, sometimes linking them by color or by retaining their positions. Thom and Ertl [2015] use word clouds to characterize hierarchical document clusters during scatter-gather-based refinement of retrieval results. Collins et al. [2009c] combine the idea of word clouds and parallel coordinates to allow for a direct comparison of term frequencies for different metadata attributes of the documents, such as time or location. While these visualizations can be used to illustrate the evolution of words in different text documents, the documents themselves are not distinguished in the word clouds. An exception is presented by Viegas et al. [2007], who show words from more than one text in a single word cloud, using font color to indicate the source of each word. This leads to words that appear multiple times. Similar to our ConcentriCloud extension, Burch et al. [2014] show words from multiple text sources in a circular word cloud, with the individual text documents at the perimeter. In contrast to our ConcentriCloud approach, an efficient use of screen space is computationally expensive with their word placement algorithm, and would require some additional heuristic strategies.

3.3 Word Clouds for Interactive Exploration

This section describes the approach based on a screen shot of our prototypical implementation. Figure 3.1 shows the Word Cloud Explorer with the Sherlock Holmes novel “The Hound of the Baskervilles” by Arthur Conan Doyle. The system consists of the central word cloud view and a number of additional components that provide further information and functionality. The individual components are

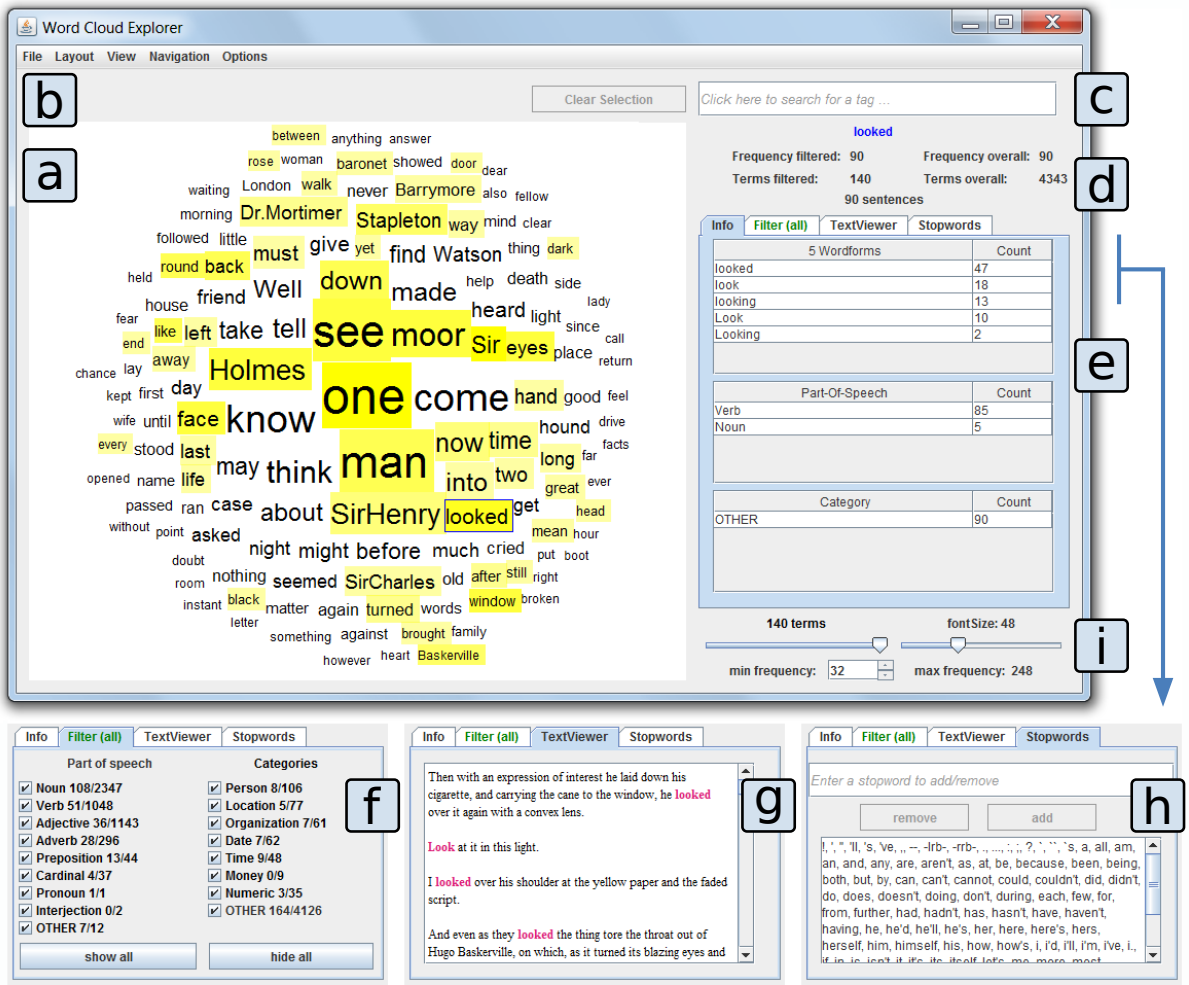


Figure 3.1 — The Word Cloud Explorer consists of the following components: (a) central word cloud view, (b) term filter, (c) search box, (d) term statistics panel, (e) info panel, (f) part-of-speech and named entity filters, (g) text viewer, (h) stop word editor, and (i) cloud control panel.

marked with letters in Figure 3.1. In the following, we describe their functionality and explain how they support users with text exploration.

After a text file has been loaded, the system performs a linguistic analysis of its contents. We use the Stanford CoreNLP tools [Manning et al., 2014] for this purpose and perform several processing steps. This includes tokenization, sentence splitting, and lemmatization. Then, based on the above-mentioned requirements that the output of the NLP facilities have to be interpretable by casual users, and the accuracy of the algorithms have to be very high, we have decided to use two

methods: part-of-speech tagging (POS), and named entity recognition (NER). Both the POS tagger [Toutanova et al., 2003], and the NER method [Finkel et al., 2005] within CoreNLP exhibit high accuracy on test corpora with newspaper style texts of greater then .95 and greater than .85, respectively. To facilitate interpretation of words in the cloud by human users, we further detect nominal multi-words and show them as one expression. For this, we have implemented a scheme that is based on the results of the part-of-speech tagger. It joins all continuous sequences of proper nouns that occur in the same sentence. With this simple heuristic, we can detect most compound nominals and proper names in the text (see Sag et al. [2002] for a comprehensive summary of different multi-word phenomena). The separate display of multi-word expressions is particularly important for many person or place names, which are often multi-words (e.g. "Michael Jordan" or "New York"). Another benefit of identifying multi-words is that the frequency counts of the individual terms are not artificially increased by fragments of multi-words (e.g. "new" as part of "New York").

3.3.1 Visual Interface

The word cloud view (Figure 3.1 (a)) implements three different word cloud layouts that the users can choose from. Two sequential line-by-line layouts, one ordered alphabetically, the other by frequency. In addition, a circular layout showing the terms with the highest frequency at the center of the cloud and the lower frequency terms at its perimeter (see Figure 3.1). The alphabetical layout supports users in quickly spotting specific terms they are looking for, while the frequency-ordered layout emphasizes high-frequency terms. Both layouts are complemented by the circular one as a space-efficient and visually appealing alternative that also stresses high-frequency terms [Lohmann et al., 2009]. Font size is scaled linearly with term frequency for all layouts. The prototype allows for an easy addition of further word cloud layouts and mapping functions from term frequency to font size of the terms. The word cloud view uses information about different word forms provided by the lemmatization component to merge them under one representative term in the word cloud. For example, all inflections of a verb are thus merged. Their counts are added up and the most frequent form is selected as representative. Detected multi-words are displayed in camel case to make them easily recognizable as one entity. As such they cannot be confused with two independent words of similar size. Both features can be disabled in the prototype, if users wish to do so. This might be beneficial for texts, e.g., in an unsupported language, that cannot be accurately processed by the NLP tools.

During exploration, users can hover over terms to highlight related ones. According to our concept, we consider two terms related if they co-occur within the same

sentence. This *co-occurrence highlighting* [Lohmann et al., 2012], that is akin to the visual concept of *weighted brushing* [Dörk et al., 2008], intuitively conveys term relations without introducing visual clutter. In our approach, related terms are marked with a yellow box whose saturation corresponds to the relative co-occurrence frequency. We have chosen this highlighter metaphor as it is very intuitive and provides an effective means to assess the 'strength' of term relations. The second effect of hovering over a term is that further information about it is displayed in the term statistics panel (Figure 3.1 (d)) and info panel (Figure 3.1 (e)). The term statistics panel (Figure 3.1 (d)) displays information about a focused term, as illustrated for the term *looked* in Figure 3.1. The statistics panel lists the number of occurrences of the term within the filtered set of sentences (*Frequency filtered*), the number of terms currently present in the word cloud (*Terms filtered*), the number of occurrences of the selected term in the whole text corpus (*Frequency overall*), and the overall number of terms in the corpus (*Terms overall*). Finally, it gives the total number of sentences in which the focused term occurs, which is identical to the above values in the depicted case, as no filters are selected. The information from the term statistics panel helps, for instance, for sanity checks. One disadvantage of word clouds is that the difference in frequency between terms gauged according to their font size can lead to a false impression about their true frequency ratio (see Alexander et al. [2016]). Showing the absolute frequency values to users lets them easily identify and correct false impressions. Under the term statistics panel, as part of the tabbed pane, the info panel (Figure 3.1 (e)) displays linguistic information about the focused term. This comprises all word forms present in the text, part-of-speech tags, named entity types, and respective frequency counts. In Figure 3.1, for example, five different word forms have been detected for the focused term. The word form "looked" is chosen as the representative because it appears most often in the text. Furthermore, the term is mostly used as a verb and is not part of any named entity (indicated by the category name "OTHER", see below for more information). This helps users during their exploration of the text body in two respects. First, it can be used to learn more about a term in the word cloud. For instance, it might be interesting to see whether a term occurs mainly in present or past tense, or if it is used as verb or noun. Second, it can be used to cross-check the results of the linguistic analysis. Although the accuracy of the text processing techniques is generally high, there are occasional errors depending on the type and quality of the text.

By clicking a term, users select it and add it to the term filter (Figure 3.1 (b)). The word cloud view shows only terms that co-occur with *all* of the selected terms, i.e. it changes from a word cloud for the whole text to a 'co-occurrence cloud' as soon as terms are selected. Figure 3.2 shows such a co-occurrence cloud for a text corpus with abstracts of visualization research [Isenberg et al., 2016]. The

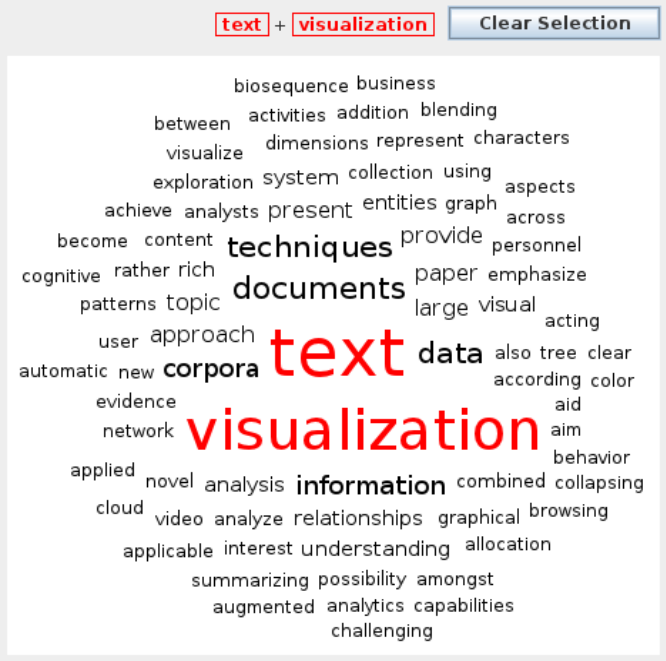


Figure 3.2 — The terms *text* and *visualization* have been added to the term filter and are highlighted in the resulting co-occurrence cloud for the VIS abstracts data set.

selected terms are colored red (in this case, "text" and "visualization"). They can be added and removed from the term filter in any order and at any time. Through this filter functionality users can focus on the co-occurrences of one term only. Thus, an effective drill down to relevant information is supported that facilitates sensemaking processes (see Chapter 2). To search for terms that users become interested in during the exploration of the data set, terms can be added to the filter with the search box (Figure 3.1 (c)). If a word is entered that occurs in the text, the term statistics panel (Figure 3.1 (d)) and the info panel (Figure 3.1 (e)) display information about it. If it is part of the word cloud, it is highlighted, along with all co-occurring terms. The search box thus allows to view co-occurrence clouds for terms whose frequency is too low to be displayed in the initial word cloud.

The filter tab (Figure 3.1 (f)) enables users to explore the word cloud according to different parts-of-speech and named entities. To keep the interface easily interpretable by casual users, we call named entities *term categories*. In addition, we condensed the quite fine-grained POS categories to nine major ones. Users can hover over POS and NE categories to highlight all corresponding terms in the cloud. Again, we use a yellow box whose saturation indicates what fraction of a term's occurrences have been tagged with the respective category. The two

Figure 3.3 — A word cloud for the text "The Hound of the Baskervilles" with the terms colored according to their most frequent part-of-speech (noun, verb, adjective, adverb, preposition, other).

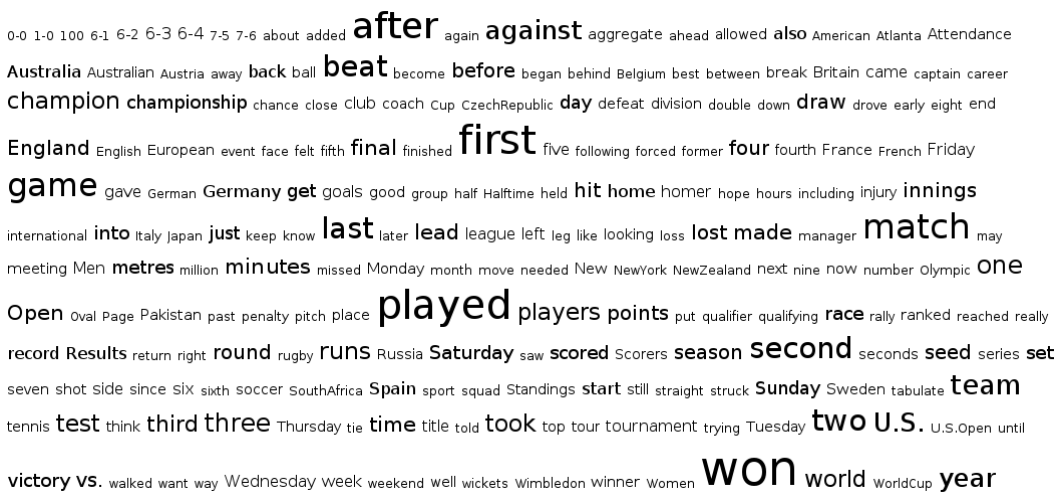


Figure 3.4 — The initial word cloud with alphabetical layout for one week of Reuters sports news.

numbers after each category name indicate the number of terms of the respective category that are part of the cloud and of the overall text set. Disabling one POS or NE category in the filter tab subtracts its counts from the overall frequency of the terms. Being able to filter by POS and NE is a powerful feature of the Word Cloud Explorer that enables users to explore specific aspects of a text. The word cloud can, for instance, be set up to show only locations that occur in conjunction

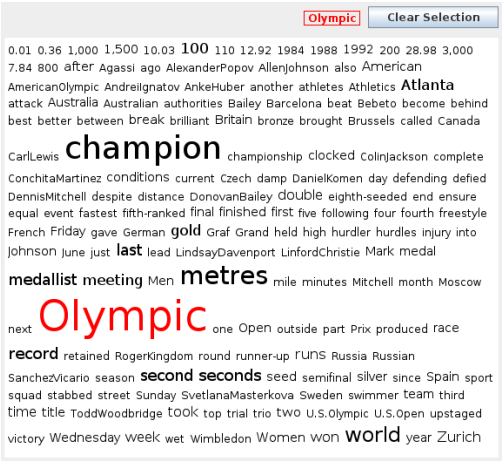
with a certain person. Or adjectives that co-occur with a specific organization can be explored.

The text viewer (Figure 3.1 (g)) lists all sentences that contain the selected and/or focused terms (see Figure 3.1 for the term *looked*). Showing the original contexts allows users to disambiguate words or refine their original ideas about their meaning within the text body. In addition, further knowledge, e.g., about connections between terms, such as the links between two persons can be acquired. Based on the text type and exploration goals, the stop word editor (Figure 3.1 (h)) allows users to modify the stop word list. This might be important for domain-specific high frequency terms that clutter the word clouds of a specific data set. The Word Cloud Explorer is highly customizable with the cloud control panel (Figure 3.1 (i)) that allows to dynamically change the size of the word cloud and of the displayed terms. Users can set the maximum number of terms, the maximum font size of terms, or the minimum frequency that terms require to appear in the word cloud. When one of these values is changed, the other values are adapted accordingly. Users can further disable stop word filtering, define a cutoff frequency for the co-occurrence calculations, disable the merging of multi-word expressions, and turn the lemmatizer on and off. The latter has the effect that different word forms are no longer merged. Another feature offered in the menu is a coloring of the terms according to their most frequent part-of-speech. A screen shot of a word cloud with this functionality activated is depicted in Figure 3.3.

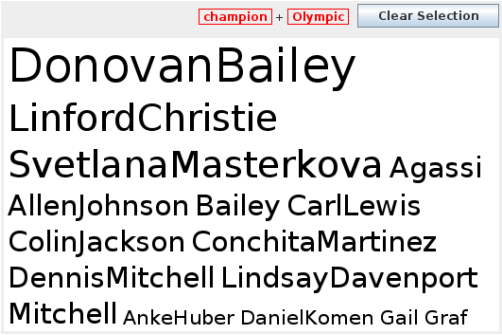
3.4 Application Example

In the following, we present an application example of the Word Cloud Explorer. The corpus we use for this is Reuters Corpus Volume 1 (RCV1) [Lewis et al., 2004]. It consists of a large collection of manually categorized news articles made available by Reuters Ltd. for research purposes. We use the first week of articles from the corpus, ranging from August 20 to August 26 of the year 1996. Three major global sports events were dominating the news during that week. First, the summer Olympic games that took place in Atlanta, GA that year. Second, the Wimbledon Tennis Championships in London a little earlier. Third, the U.S. Open, a tennis championship, that started in New York. We therefore restrict the Reuters corpus to sports news by selecting all articles that have been categorized accordingly. The task for this example is the open exploration of the data set to learn about the themes and information contained within the data set.

To get a first overview, we choose an alphabetically ordered word cloud as shown in Figure 3.4. The most frequent terms in this cloud are common expressions from the sports domain, such as *won*, *played*, *match*, and *game*. We can also spot some



(a) The alphabetically ordered co-occurrence cloud for the term *Olympic*.



(b) The frequency ordered co-occurrence cloud for the terms *Olympic* and *champion* filtered by person names.

(c) The frequency ordered co-occurrence cloud for the terms *Olympic*, *champion*, and *Svetlana Masterkova*.

Figure 3.5 — Word clouds of the application example.

names of countries, cities, and sports events, such as the aforementioned *Wimbledon*, *Olympic* games, and *U.S. Open*. From this, we are instantly able to learn about the coarse themes of the data set. To dig a little deeper below the surface and be able to gauge the information depth of the data set, we select the term *Olympic* to learn more about it. In the term statistics panel and the info panel, we can see that it occurs 90 times in 87 sentences, seven of them are occurrences within multi-words such as *Olympic Committee*. We switch to the co-occurrence cloud of *Olympic* (Figure 3.5a) and see that the term *champion* is used most often with it. Among the other terms in the cloud, there are many athletes with their first and last names merged by the multi-word feature. We further see numbers denoting scores, years, distances, etc. This indicates that lists of winners and result tables are part of the data set. To learn more about the level of detail contained in the data set, we add *champion* to the term filter and choose the frequency based ordering. We then use the named entity filter to show only persons in the cloud and aim to review the information accessible for a specific example (Figure 3.5b). We concentrate on *Donovan Bailey*, the most frequently mentioned Olympic champion. To read more about him, we open the text viewer that lists all sentences containing his name. We learn that he is a Canadian sprinter, and get the information that he set a speed record at the 1996 Olympic games. In addition, we can find similar detailed information for the Russian athlete *Svetlana Masterkova* (see Figure 3.5c).

The application example showcases the applicability of our word cloud based text exploration system that facilitates learning about a text data set. Initially, users quickly get a rough idea of a text's themes. The depth of information can then be explored and gauged through systematic drill-down analysis that is supported down to the single text snippets. All the interaction and filter functionality of the Word Cloud Explorer, including the NLP techniques, support and speed up these exploratory analysis processes.

3.5 Initial Feedback and Discussion

To learn about the applicability of the approach, we solicited initial user feedback from five users, which were all members of our visualization institute (25-31 years of age). For this, we used all of the three corpora presented above. The participants were handed questionnaires during the sessions that contained example tasks they were asked to solve with the Word Cloud Explorer. In the context of this thesis, we concentrate especially on the discussion of those tasks that comprise exploratory elements. A more extensive discussion can be found in Heimerl et al. [2014]. Before we discuss the procedure and results of the user sessions, we want to point out the limits of this initial user feedback. Although visualization experts are not the users that we primarily target with this approach, we decided to rely on them for

two reasons. Firstly, due to limited resources we were only able to accommodate a rather small number of users. And secondly, experts in interactive systems are able to give profound feedback and find problems with the approach faster than the average user due to their experience with such systems. The drawback of this decision is that the results have a technical bias towards visual encodings, interaction, and implementation quirks. Nevertheless, we think that the discussion of the results demonstrated the applicability of the Word Cloud Explorer, and provides valuable feedback with respect to its design and implementation.

3.5.1 Procedure

The procedure for each of the participants consisted of the following five steps: *i)* color vision deficiency test with the Ishihara color plates. *ii)* Brief user introduction with the "The Hound of the Baskervilles" corpus. The participants could ask questions and try out the system until they felt confident to use it. *iii)* Completion of tasks on paper based on the Reuters and the VIS abstracts corpus. In case participants were stuck during a task, we kept hints on a separate sheet of paper that they could consult. To get additional insight, we encouraged the participants to articulate their thoughts during task completion according to the think-aloud method. *iv)* The participants filled in a questionnaire with questions about their background and thoughts on the approach. *v)* Finally, we asked the participants for any additional feedback.

3.5.2 Results and Discussion

The participants rated the Word Cloud Explorer as an intuitive and useful text analysis system. They were impressed by the wealth of possibilities that such a straightforward visualization paradigm enriched with context information, filters, and interaction offers, and they were able to quickly solve all the tasks listed. This was especially true for exploration-based tasks, such as "Who are important persons for the 1996 Summer Olympics" for the Reuters data set. However, some users remarked that they would only use the approach in combination with other tools complementing its functionality due to some missing capabilities. These are mostly of an analytical nature, such as the limited search functionality of the text viewer, and did not affect the explorative functionality of the prototype. With respect to the word cloud layouts, an interesting finding was that all participants preferred the sequential layouts over the circular one, although they rated the circular one to be aesthetically most appealing. When asked about this apparent contradiction, most participants answered that they found it easier to visually compare relative word sizes using the line-by-line layout. This is because the lines could be used as visual anchors which allow to compare font height. Furthermore, we could observe the

participants switching between the frequency and alphabetically ordered layouts according to whether they were interested in high frequency terms or searching for a specific term. This indicates that it is important to provide different word cloud layouts that users can choose from depending on the analysis task. The participants were in disagreement about the usefulness of the part-of-speech coloring function. Some considered it a useful feature, while others found that it has little analytical value. It was argued that the part-of-speech of most words is known by users once they read the word and does not have to be marked. This outcome was to be expected and illustrates the language skills of human users. Using the part-of-speech categories as a filter for the word cloud, however, was considered useful. Through automatic linguistic analysis, it saves users from reading large amounts of texts and filters and aggregates it efficiently. For the same reasons, the named entity feature was unanimously found helpful and the participants used it frequently to solve the tasks. The aggregation of multi-words and different word forms was also mentioned positively. Overall, the participants assigned many positive attributes to the word clouds, such as "tidy", "clear", "efficient", "useful". To summarize, this initial used feedback provides some indications that the linguistically and interactively improved word clouds are indeed an adequate and effective means for exploratory analysis of single text bodies.

3.6 Extension to Multiple Documents

The Word Cloud Explorer is capable of processing multiple documents, such as series of novels for literary analysis. Through the text viewer, single terms can also be tracked back to their original documents. Exploring the contrasts of term occurrences between multiple documents, however, is not supported by the Word Cloud Explorer concept, as the documents that each term in the cloud occurs in are not directly discernible. However, such exploration and analysis problems are also frequent in casual analysis scenarios, e.g., when comparing novels of a series, or different dissertations from one university department. For this reason, we have extended the previous concept to an approach for multiple documents that solves these problems. In this section, we present and discuss the concept that we call ConcentriCloud, including its prototypical implementation.

3.6.1 Concept and Example

ConcentriCloud is composed of several word clouds representing different combinations of the text documents. This is akin to the small multiples concept, because it includes comparable views of the document terms. However, through systematically merging the word clouds, commonalities and differences between

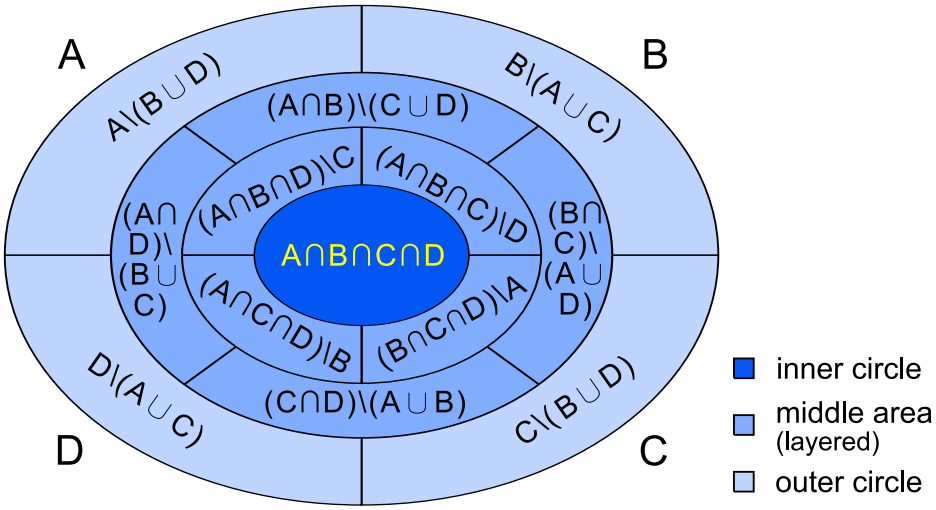


Figure 3.6 — Schematic illustration of the composition of ConcentriCloud (letters A to D represent the bags of words of four text documents).

the documents can be more easily identified. Furthermore, ConcentriCloud avoids redundancies, as terms are usually displayed only once, with the exception of some special cases that are discussed later. ConcentriCloud arranges the word clouds on concentric circles, as sketched in Figure 3.6 for four documents A,B,C, and D. Each document is represented by a set of words that comprises the terms from the document along with their frequencies. Formally, each document can be defined as a set of terms $T := \{t_1, \ldots, t_n\}$ with individual terms t_i, $1 \leq i \leq n$. Each term is additionally associated with its frequency value. The sets of terms T_x are combined into several word clouds W_y, each representing a different combination of the documents. The word clouds on the outermost circle contain terms that occur in either of the documents, with some notable exception, as discussed below. For instance, the word cloud representing document A consists of terms that are contained in A but not in B and D (i.e., $W_A = A \setminus (B \cup D)$). Document C is an exception in this example, as its word cloud is located at the opposite side of the outer circle. Such word clouds are not merged on any layer of the middle area, as there is no intuitive position for such a composite cloud in the concentric layout, except from the inner circle. The inner circle, however, is reserved for words that occur in all documents and not only in a specific subset of documents. This results in the fact that terms can appear more than once in the visualization if two opposite word clouds contain the same term. However, as we found out through user evaluation [Lohmann et al., 2015], this redundancy due to layout constraints only marginally affects the general readability and interpretation of the visualization. This is especially true if it is visually indicated, for instance, by

interactive highlighting. In each of the word clouds towards the center, the terms from the documents are systematically combined, i.e., the layers on the middle area contain terms that occur in more than one document (but *not* in all documents). For instance, the second level of the circle contains word clouds that represent pairwise intersections of the documents minus the pairwise unions of the rest of the documents, with the aforementioned exception that oppositely located word clouds are not combined. In case of documents A and B, this results in a word cloud $W_{A,B} = (A \cap B) \setminus (C \cup D)$, among others (see Figure 3.6). Finally, the innermost circle consists of only one word cloud containing those terms that occur in *all* documents. In the illustrated case, it thus represents the intersection of all four documents, i.e., $W_{A,B,C,D} = (A \cap B \cap C \cap D)$. For any other number of documents, the composition of the visualization needs to be adapted accordingly. As a general rule, each ConcentriCloud is theoretically composed of as many circles as there are documents. Since this can result in a large number of circles, certain layers in the middle area may be skipped, as long as the overall composition principle remains the same. However, note that terms should always appear on the highest possible aggregation level in ConcentriCloud, i.e., on the level closest to the center. The only exception are terms from oppositely located word clouds, or, more generally, from word clouds that are not neighbors on the outer circle. In case of three documents, there is no such exception, but with an increasing number of documents, the likelihood of term redundancy increases. One strategy to minimize any remaining term redundancy in the visualization is to order the documents based on content similarity, as we have done in our prototype.

Figure 3.7 shows an example of ConcentriCloud that visualizes frequent terms of all seven "Harry Potter" novels. The word clouds on the outermost circle represent the individual novels (HP1 to HP7). They are visually separated by lines, while the names of the source files are shown next to them. Examples of terms that appear in only one of the novels are *lockhart* (second novel) and *karkaroff* (fourth novel). The angular size of the word clouds indicates the relative length of each novel, which increases for the Harry Potter novels. Terms that can be found in all seven novels are shown in the inner circle of the visualization, such as *harry* or *dumbledore*. Since the inner and outer circle are most important for the idea of ConcentriCloud, because they facilitate contrastive exploration of texts, they are clearly distinguishable in the visualization. Borders between the layers in the middle area are omitted to produce a clearer picture and to reduce visual clutter that would be introduced by too many separating lines. If a word cloud in the middle area does not require all the reserved screen space, it is used by neighboring word clouds to place further terms beyond their bounding box for a more space-filling design. However, the general composition principle remains the same, i.e., the closer a term to the center, the more documents contain it. This

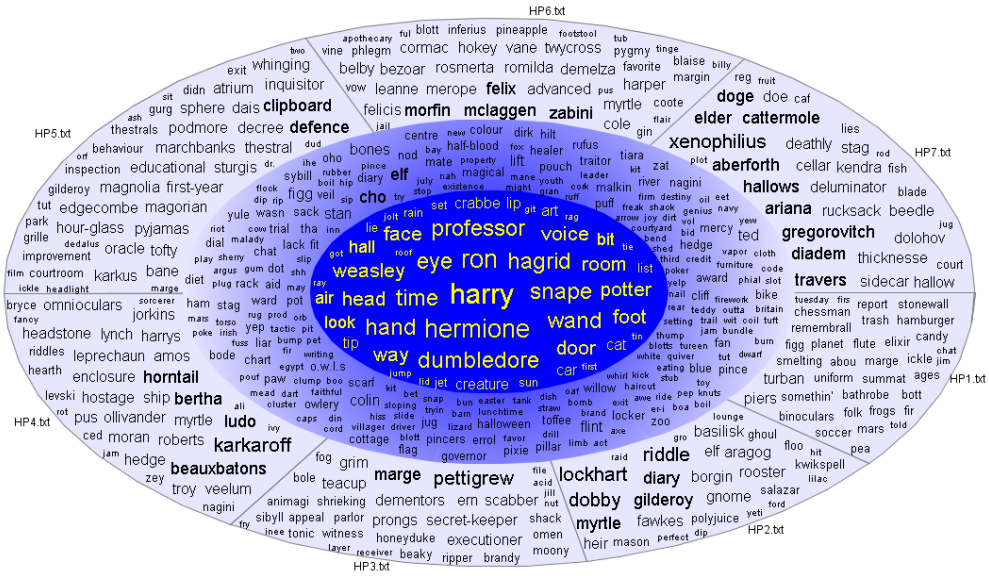


Figure 3.7 — ConcentriCloud visualization of all seven *Harry Potter* novels (HP1 to HP7).

principle is additionally emphasized by the saturation of the background color, which has a gradient towards the center in the middle area.

3.6.2 Design and Implementation

We implemented the approach in a stand-alone Java-based prototype that generates multi-document word clouds. Currently, it is not integrated into the Word Cloud Explorer, but serves a proof-of-concept for the support of contrastive exploration of documents based on the word cloud concept. In the following, we describe our design decisions and implementations-specific details.

The first step when creating a ConcentriCloud is to process the documents and extract meaningful terms and their frequencies. For this task, we again use Stanford CoreNLP [Manning et al., 2014], in particular, its tokenization, lemmatization, and part-of-speech tagging features. The part-of-speech information can be used to create word clouds that include only nouns, such as the Harry Potter word cloud in Figure 3.7, while the lemmatization is used to produce cleaner word clouds by merging morphological term variants. We transform all terms to lowercase, and apply a stop word list to the extracted lemmas to remove high-frequency terms.

There are several possibilities of arranging the documents on the outermost circle. We have implemented two different schemes. The first allows for a manual ordering by the user, which may reproduce some natural ordering of the documents, e.g., according to their publication date. An example for this are the Harry Potter novels in Figure 3.7, ordered from the earliest to the latest publication in this series. As a second way of arranging the documents, we developed an algorithm that orders them according to their similarities. After computing the cosine similarity for each pair of documents, the algorithm greedily chooses the highest similarity score between two documents and reduces the set of possible orderings to those in which both documents are neighbors. It continues recursively until each document has a fixed position. This similarity-based ordering was used to create the cloud shown in Figure 3.8. After the ordering of the documents, the size of the word clouds on the outer circle is determined by scaling the angle according to the document length. We implemented a scheme that is able to accommodate the worst case scenario in which one very long document will take all the space from several very short documents. This is solved by splitting the available 360° into two parts, assigning each of the n documents a minimum angle of $\frac{180°}{n}$, and allotting the remaining 180° degrees according to document size.

ConcentriCloud attempts to render the terms within the bounding box of the respective word clouds between the concentric circles. Similar to Burch et al. [2013], the terms are placed along invisible concentric circles from the center of the respective word cloud, starting with the most frequent term and continuing with terms of decreasing frequencies. This ensures that the most frequent terms are placed first and that they appear in the center of a cloud, such as the term *harry* in the inner cloud of Figure 3.7. If a term cannot be placed within the cloud's bounding box, it is skipped, and placing the next term from the frequency-ordered list is attempted. This placement strategy has the limitation that some high-frequency terms may not be rendered due to their larger size and space limitations, whereas other low-frequency terms of smaller size could be placed in the word cloud, as they fit in the available space. However, this is rather a general limitation of word clouds than a particular drawback of ConcentriCloud. An alternative strategy would be to stop the placement and to not add smaller terms in the available space, as soon as a larger term cannot be placed. Yet, this could result in a lot of unused space that may better be filled with smaller terms, also for aesthetic reasons. The font size of the terms is scaled either linearly or logarithmically with their occurrence frequency, depending on user selection. If a word cloud represents the terms from more than one document, we use the average term frequencies to scale the font size.

Our implementation does not only create a static word cloud, but it includes options to customize and interact with the visualization. Before the visualization is created,

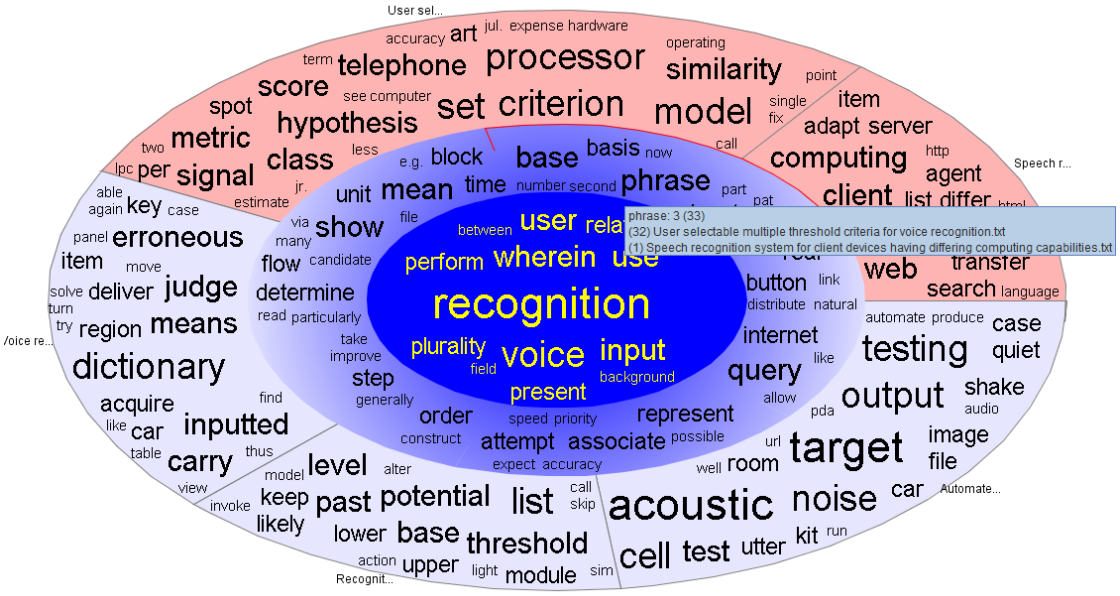


Figure 3.8 — ConcentriCloud visualization of five patents on the topic *voice recognition*. The term *phrase* is hovered and related patents are highlighted in red.

users can specify the ordering of documents, and term filters, e.g., by part-of-speech. While Figure 3.7 was an example of a noun-only ConcentriCloud, Figure 3.8 displays also other parts-of-speech, such as verbs and adjectives. Figure 3.8 shows two additional interaction possibilities: (1) highlighting of the word clouds that represent the corresponding documents when the user hovers a term in (in this case *phrase*), and (2) tool tips that appear for each word showing its overall number of occurrences in the document set and its distribution across individual documents. These interactions help users get a better understanding of the word cloud composition and exact term frequencies.

3.6.3 Limitations

Clearly, there are limitations to this concept. Although technically, ConcentriCloud can scale up to an arbitrary number of documents, there is a limit to visual scalability at around a dozen documents. It is thus applicable to small sets of consecutive documents, such as series of novels in literary analysis, or sets of publications from the same authors or about similar topics in scientific literature analysis. Moreover, it is important to note that word clouds usually do not show all terms of a text document but only the most frequent ones. Due to the space-filling

layout, smaller words may be added to the word cloud if larger ones do not fit in the remaining screen space. In order to avoid inaccurate interpretations in such cases, we provide a list of all terms and term frequencies for each of the word clouds on demand. In this context, anchors back to the original text source would also important. In addition, there are some phenomena that can appear when using ConcentriCloud. For instance, if the analyzed documents do not have any common words, the inner circles would be empty and only the word clouds on the outermost circle would display terms. In the opposite case, if documents are identical, the outer circle would be empty and terms would only appear in the word clouds of the inner circles. Although such extreme cases are rather unlikely, they illustrate the limitations of the approach and indicate that it may not work equally well in all situations. While for many small sets of documents, ConcentriCloud provides a lucid depiction to support the exploration of similarities and differences between single texts, for larger, more heterogeneous data sets, different visual exploration approaches are better suited, such as the one we discuss in the following chapter (Chapter 4) of this thesis.

3.7 Future Directions

In this chapter, we have presented a highly adaptable and versatile approach for casual text exploration and analysis tasks. With it, we explore the combination of the word cloud visualizations and high-accuracy NLP methods, namely POS tagging and NER, to support exploratory analysis of single text bodies. As proof of concept, we developed the Word Cloud Explorer, a prototypical system that uses word clouds as its central visualization method and integrates several interactive features into one consistent framework for text exploration. We demonstrated the applicability of the approach with a usage example and provide first insights into its effectiveness based on qualitative user feedback. In addition, we extended the word cloud approach to support the contrastive exploratory analysis of multiple documents. ConcentriCloud, the extended version of the approach, provides users with a first impression of word use in the different documents and supports the visual identification of differences and commonalities. We implemented a separate prototype that exemplifies this concept equipped with multiple interaction techniques that provide details on demand.

We see possible future research endeavors along three lines. Firstly, the user feedback provided first insights into the applicability of the Word Cloud Explorer approach, but it is still very limited. For this reason, we aim at a broader evaluation, of both approaches. The most effective way to go about this would be to combine both approaches into a web-based implementation that allows users to upload their own documents for exploration. This could be combined with a questionnaire, and

in addition, statistics about the data sets could be collected. Thus feedback from a wide range of casual users could be collected and evaluated. Secondly, we see another area of application for the ConcentriCloud approach. So far, we have applied it to series of documents, but they could also serve as a comparative visualization for different topics extracted from a large text collection by a topic modeling algorithm. Here, too, intersections and unions of set of weighted terms have to be visualized. The multi-level approach could provide a deeper insight into similarities and differences between topics than current matrix-based techniques [Chuang et al., 2012b; Alexander et al., 2014]. Thirdly, both the Word Cloud Explorer and the extended concept could be integrated into a text analysis system that facilitates the exploration of large document sets, such as the one discussed in the following Chapter 4. Here, users could be supported with closely analyzing locally confined, small amounts of documents that they have identified as particularly interesting. Based on this approach, they could learn about the breadth and depth of information contained therein as part of a larger, exploratory analysis session.

Chapter 4

Exploratory Analysis of Document Spatializations

This chapter contains previously published material from the following publications:

F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl. Docucompass: Effective exploration of document landscapes. In *Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Oct 2016b

Although the previously discussed approach supports exploration of multiple documents, the sizes of the data sets it supports are clearly limited to around a dozen, despite the ConcentriCloud extension that we introduced for comparative exploration and analysis. This chapter discusses a novel interaction technique that is applicable to larger data sets of documents, and presents possible extensions and adaptions. The document sets can be huge in size and heterogeneous, as the technique exhibits a high scalability according to both aspects. In addition, our technique is very adaptable. Its configuration can be modified and adjusted to fit data set type, size, and the goals of exploration depending on the requirements of a given approach or system.

4.1 Motivation

A popular visualization method for large text collections is to represent each document with a glyph in 2D space. These landscapes can be the result of optimizing pairwise distances in 2D to represent document similarities. Alternatively, the spatializations are provided directly as meta data of the documents, such as geo-locations. For well-defined information needs [Hearst, 2009], analysis goals, or extraction tasks that are known in advance, suitable interaction methods are available for such spatializations. These methods require users' previous knowledge about the contents of texts and the information they are looking for. However, support for free exploration and navigation on a level of abstraction between a labeled document spatialization and reading single documents is largely missing. This limits the usefulness of document spatialization approaches for exploratory analyses during which human sensemaking processes play an important role (see Chapter 2). To fill in this gap, we created the DocuCompass technique, a focus+context method based on the lens metaphor. It comprises multiple methods to characterize local groups of documents, and to efficiently guide exploration based on users' requirements. DocuCompass thus facilitates effective interactive exploration of document landscapes without disrupting the mental map of users by changing the layout itself. It works with 2D document spatializations [Wise et al., 1995], and supports interactive abstraction and explication tasks on subsets of text documents. Both coarse and fine-grained exploration down to the level of single documents is supported through interaction. Moreover, our magic lens-based technique is very flexible and extensible by many different characterization methods for document sets. According to Cockburn et al. [2009], it can be considered a cue-based focus+context technique, depending on the configuration (discussed in Section 4.3). Through the large number of configuration possibilities, different levels of exploration, text types, and analysis scenarios can be addressed to help users develop and evolve their information needs during exploration. We used lens-based interaction approaches for text exploration as a useful part of VA tasks and systems in previous work [Bosch et al., 2011; Heimerl et al., 2012b; Bosch et al., 2013]. However, a comprehensive discussion of the interaction design and configurations for various purposes is not available yet. With DocuCompass, we improve previous approaches into several directions and provide the following improvements to the current state-of-the-art:

- DocuCompass constitutes an advancement for text exploration tasks by facilitating explorative analyses of large text collections. In addition, it offers navigation support to help users form and solidify an information need during initial, explorative analysis.

- This chapter discusses the lessons we learned by applying differently configured text lenses on various types of text spatializations as part of previously presented VA approaches.
- We extend existing versions systematically, and discuss the design space of visual document characterizations shown with the lens, text extraction and analysis, and possibilities for supporting users with navigation cues on different types of texts and spatializations.
- Finally, the results of a preliminary user study are presented that indicates the effectiveness of our technique.

Although many factors play a role for creating useful approaches, a magic lens-based technique has the benefit of keeping the context of the lens visually unchanged, or at least static with respect to the geometrical position of visual elements. From our perspective, lens-based techniques are therefore a natural fit for exploring text collections. They can be used on a wide range of different spatializations, and enable dynamic filtering, visual enhancement, and, as we exemplify in this work, even interactive data mining on a subset of a text corpus freely chosen by users. We see DocuCompass as a powerful interaction approach to complement traditional techniques, which offer either overview or details and lack intermediate interaction methods.

4.2 Related Work

DocuCompass offers an effective means to explore 2D *spatializations of text collections*. It is a *lens-based focus+context interaction* approach that aims to expedite and improve exploration of text spatializations to support knowledge generation through human sensemaking.

4.2.1 Spatialization of Texts

A straightforward way to lay out documents is with an intrinsic 2D mapping, such as geo-locations [MacEachren et al., 2011]. Alternative methods are either based on other metadata or content. Zhao et al. [2013] allow users to create and compare scatter plots according to different attributes of articles in a scientific data set and displays citation links between the documents. Galaxies [Wise et al., 1995] are a 2D point cloud of documents, in which proximity indicates content similarity. An extension, Themescapes, are 3D density plots based on a topographical map metaphor to depict topic peaks in 2D document space. Both approaches are based on the vector space model, and map this high-dimensional space into 2D, preserving

pairwise distances as well as possible. Since then, the idea has been used widely. This includes commercial packages such as IN-SPIRE™ [Wise et al., 1995] and Aureka[1]. Correll et al. [2011] use spatializations in an approach to explore collections of tagged texts. Mapping high-dimensional vectors into 2D introduces errors with respect to the pairwise distances. Recent approaches to reduce these errors are least squares projection (LSP) [Paulovich et al., 2008] and t-distributed stochastic neighbor embedding (t-SNE) [van der Maaten and Hinton, 2008]. Although these methods improve projection quality, they still result in information loss. This is particularly problematic during exploration tasks, as it can lead to misjudgments of document similarities. Another challenge is the characterization of the resulting 2D space, allowing users to understand its organization. For this, approaches based on identifying coherent regions of the space and finding representative labels have been proposed [da Silva et al., 2015; Kandogan, 2012]. Though this concept has been applied to high-dimensional text data [Choo et al., 2013; Wise et al., 1995], a very large number of dimensions hinder the selection of useful areas and terms. Our exploration technique alleviates this problem by supporting varying granularity levels and different document set characterizations, thus flexibly adapting to users' analysis needs. There are only few spatialization approaches that allow to adapt the underlying placement model. Endert et al. [2011, 2012] let users change the position of documents in projection space, e.g., to move similar documents closer to each other. Their "semantic interaction" concept incorporates such user feedback into document placement decisions. The problem of effective exploration to understand 2D layouts, however, remains unsolved. Users still have to read single documents, or start with an initial information need formulated as a search query. Alexander et al. [2014] provide visual interactive access to text corpora on three fixed levels, the word, the document, and the corpus level. However, their approach does not allow to explore texts at arbitrary levels of granularity in between those three. Typograph [Endert et al., 2013] is close to our approach in that it supports exploration of a spatial layout of keywords on multiple levels including phrases, snippets, and entire documents. While it lays out keywords extracted from documents, our lens-based technique directly operates on document clouds. It is thus more flexible in supporting content-based layouts as well as geographic and meta-data based ones, where altering document positions is either not desired or impossible. In addition, our technique can be integrated with a wide variety of document characterization methods, based on contents and meta data.

[1] http://ip-science.thomsonreuters.com/m/pdfs/aureka_factsheet.pdf

4.2.2 Focus+Context Interaction and Lenses

Different focus+context techniques exist for various types of data [Cockburn et al., 2009], including magic lens techniques for text analysis. Around the time Furnas [1986] proposed fisheye views (an unofficial version has been available since 1981 [Card et al., 1999]) for structured information, Spence and Apperley [1982] published their approach to quickly access large amounts of text. Later, Mackinlay et al. [1991] developed a focus+context technique that provides meta data context of documents and shows textual details in the focus region. The "document lens" [Robertson and Mackinlay, 1993] is a focus+context view on larger documents. All these approaches modify the focus area geometrically or structurally to show additional details about single documents. There is no approach that shows content information of multiple documents to support explorative tasks, thus limiting scalability to the number of documents and the visual placement of glyphs. Magic lenses [Tominski et al., 2014] are a versatile [Bier et al., 1993] and straightforward way to realize the focus+context concept. Only few magic lens approaches exist to explore and navigate text collections. To the best of our knowledge, they made their first appearance in the contribution to the VAST challenge 2011 [Bosch et al., 2011]. From the exploration of geo-located micro blog messages, their use has been extended to other domains, such as disaster management based on micro blog messages [Bosch et al., 2013], and the exploration of larger documents [Heimerl et al., 2012b]. These approaches are mostly limited to document frequency to weigh and select terms that provide content information about focused documents.

We improve magic lens approaches for text according to multiple aspects. The lens technique is particularly useful during exploration. Analysis scenarios, in which users start out with an unspecified or very coarse information need, require an explorative approach that provides meaningful summarizations of the data or some of its aspects. For text documents, this has been achieved with word clouds. These contain text labels as visual elements whose optimal placement poses a problem in itself [Luboschik et al., 2008]. Word clouds have been extended in multiple directions into interactive visual interfaces for text analysis [Collins et al., 2009c,a; Liu et al., 2015b; Heimerl et al., 2014]. Depicting word clouds inside lenses likely occludes the region most interesting to users. Moving them into a separate view, however, requires users to split their attention between two regions and thus eliminates the focus+context aspect. An adequate compromise is to place labels in the immediate vicinity of the lens, but outside its focus area. Bertini et al. [2009] discuss placement variants in the context of depicting names next to focused regions, including cues for linking individual names. Lenses can also help to understand and cope with deficiencies or uncertainties of a visualization. There are approaches for analyzing and evaluating the output of projection models for text

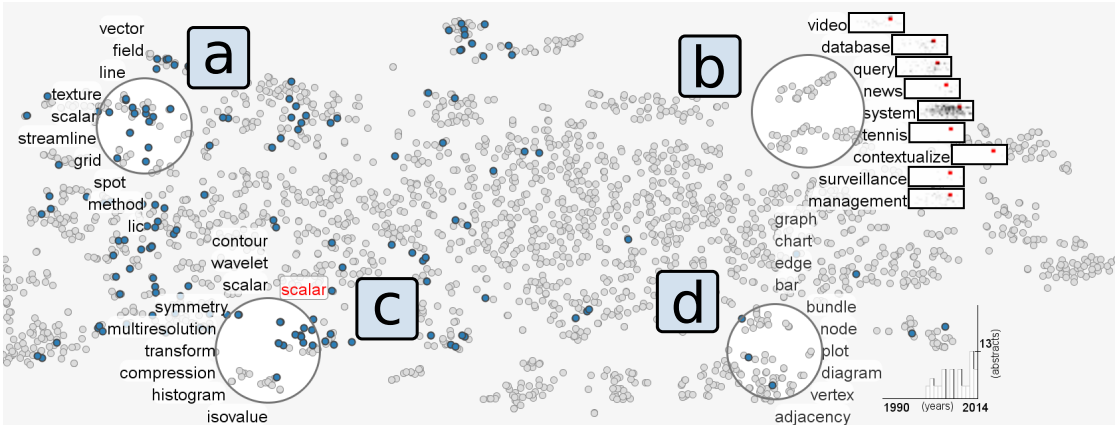


Figure 4.1 — DocuCompass comprises lenses with different features: (a) a lens showing terms as text labels for characterizing focused documents, (b) a lens depicting previews of term distributions, (c) a lens with the term 'scalar' selected, (d) a lens using a bar chart to depict the distribution of publication years.

data [Chuang et al., 2012c] and high-dimensional data in general [Schreck et al., 2010]. Stahnke et al. [2016] focus on understanding 2D layouts after dimensionality reduction, and the importance of various data dimensions. Lens-based approaches can also be used to convey information about the original, high-dimensional space to better validate projections and gauge information loss [Heulot et al., 2013]. As opposed to these approaches, our techniques focus on exploring and characterizing text documents. This is true despite the fact that we discuss methods that can be integrated with DocuCompass to reduce local ambiguity and uncertainties caused by information loss.

4.3 Exploration of Document Spatializations

We have designed DocuCompass as a flexible interaction technique that can be combined with any type of 2D document spatialization. To achieve this goal and efficiently support exploration, we have identified five design goals based on our previous experience with free exploration of document data sets:

1. Facilitate flexible analysis on arbitrary levels of granularity of the data set
2. Provide summarizations or characterizations of the focused document set that can be computed at interactive speed and that can be easily followed and quickly processed by humans

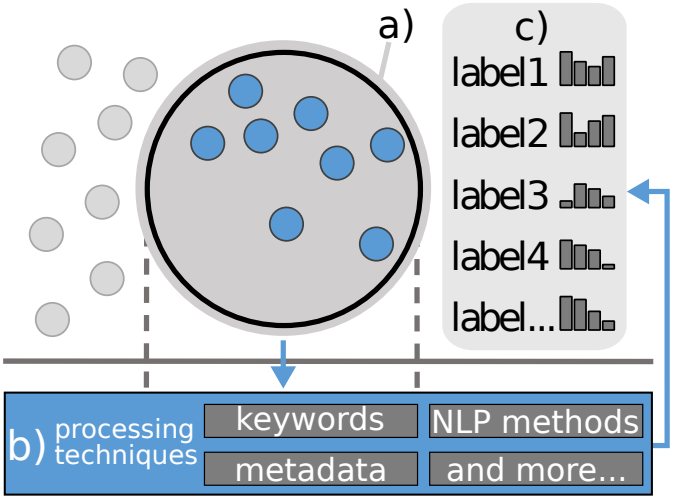


Figure 4.2 — DocuCompass includes three tightly integrated components for interactive exploration of 2D document landscapes. The first one is a flexible visual metaphor to help users specify a set of interesting documents (a). Once the set of documents is chosen, it can be further processed to extract and sort keywords. In addition, metadata aspects can be extracted, or NLP methods can be used to characterize the documents (b). A third visual component selected by users or the system then conveys these characterizations to users (c).

3. Spatial proximity of characterizations and focused document sets without covering important information to keep the exploration process efficient

4. Provide navigation support for exploration on a global and a local scale

5. Support arbitrary types of 2D spatializations and different types of document data sets

In the remainder of this section, we discuss the design of DocuCompass based on these five goals. Since DocuCompass is designed as a focus+context technique, it is tightly connected to the underlying spatialization. It consists of three building blocks (see Figure 4.2). This is, firstly, the lens to focus document sets within 2D spatializations. Secondly, a text processing method to characterize the focused document set. And thirdly, based on these text processing methods, visual characterizations of the focused texts. These are continuously updated during lens movement. DocuCompass can be flexibly configured with different text analysis techniques and a variety of visual representations.

4.3.1 Design Decisions

Users can freely move DocuCompass lenses by clicking and then dragging them. When users hover over documents, DocuCompass shows text labels, and other optional representations, such as bar charts or term distribution previews, depending on its configuration.

Placement of Cues DocuCompass displays around ten terms as text labels, sorted according to the selected weighting scheme from top to bottom next to the lens (see Figure 4.1 (a)). This lets users quickly explore certain regions of the landscape and collect impressions and insights about the contents and the diversity of the documents. In partial fulfillment of **goal 1**, we have decided to show roughly ten terms. This helps users maintain an overview and quickly recognize changes when moving the lens. To clearly separate terms, we arrange them vertically. The resulting labels' size and, accordingly, the list's height is retained during lens interaction. This makes it easier to place the terms in immediate vicinity of the lens. In accordance with **goal 3**, we have decided to place additional visual representations right next to the terms they pertain to. This has the benefit of reducing the overlap with the document spatialization, but might decrease the comparability of these optional visual representations.

We depict all visual characterizations close to the lens, but outside its focus area. This is in fulfillment of **goal 3**, since placing them inside the lens would clutter the area users are interested in. Placing cues close to the lens reduces the need to split attention between spatially disconnected regions. Bertini et al. [2009] have made a similar choice. In difference to their approach, the labels and visual representations we show do not always have a one to one correspondence between visual cues and visual items under the lens, making it impossible to directly depict links. This choice, however, means that occlusion concerns documents close to the lens, that users might be interested in (see **goal 3**). In addition, highlighting of these documents might also be occluded, for example, when selecting a term for navigation. As for labels, the problem is exacerbated by showing the label's bounding box in a color that increases contrast and thereby readability. To reduce occlusion effects, we draw a label's background semi-transparently. In addition, we show the visual characterization at the opposite side of the lens's horizontal displacement. This assumes that users have a higher interest in those regions of the spatialization towards which they are moving the lens. Users can also flip the visual characterizations to the other side of the lens on demand if the lens is not moved. Furthermore, we flip the characterizations to the other side if they would be drawn over the display's edge otherwise.

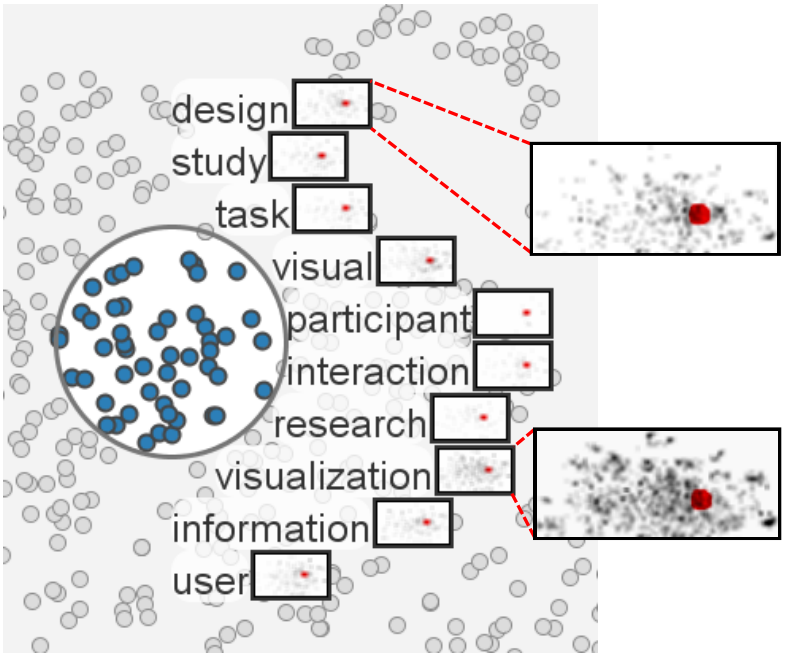


Figure 4.3 — A lens with mini heat maps to preview term distributions across the display. The red areas depict the position of the lens. As the differences between heat maps come across as rather subtle in the screen shot, we have enlarged two of them for illustration. The term "visualization" is much more prominent in the data set than the term "design", although they roughly cover the same areas. Users interested in the term "design" could start analyzing the cluster on the bottom right where the term has a particularly high prominence.

Navigation Exploring and analyzing a 2D document landscape is a challenging and complex task. In layouts generated with dimensionality reduction, exploring the space involves discovering different topics and uncovering the general structure of the document space. Exploring geographical layouts involves finding and analyzing interesting regions and learning about the type of documents and topics located there. To help users explore such spaces and guide them towards uncovering new and potentially relevant insights about the mapping and the underlying document set, we have equipped DocuCompass with advanced navigation features. These support navigation on two different scopes. Local navigation supports users with optimally placing and adjusting the lens by providing information about the internal structure of a focused area. Global navigation helps them identify and explore regions similar to previously identified documents in that they share important terms or meta data features, e.g., the same authors. This helps to extend the

analysis at hand and quickly and extensively explore all potentially relevant regions of the document space.

Global Navigation In partial fulfillment of **goal 4**, DocuCompass helps users with global navigation on the document scatter plot. Previews of a term's distribution can be displayed optionally as small multiples of the spatialization. They include a heat map that shows how frequent a term is used in other areas of the display. This helps users to quickly assess which term might be of particular interest to them. Figure 4.1 (b) depicts examples of term distribution previews. Apart from showing visual results next to the lens, documents outside the focus region can be highlighted for navigation purposes. By hovering over a term, all documents are highlighted that contain it. In addition, users can select and pin a term by clicking on it. Although this breaks to some extent with the focus+context approach, this functionality is important, because exploration often involves following a particular line of inquiry, e.g., when an interesting term catches the user's attention. The highlighting of respective document glyphs is done with pre-attentively perceptible encoding (here color) in order to reduce the time required for planning subsequent exploration tasks. This is shown in Figure 4.1 (c), with the selected term *scalar*. Once an interesting term is highlighted and all the documents that contain it are marked, the lens can be moved, or a new lens can be created. This way, regions containing documents with the relevant term can be further explored. By using the mouse wheel, users can adjust the size of the lens. This supports seamless switching between levels of various granularity. These design decisions all consider mouse interaction. If an application on a touch interface integrates DocuCompass, placement of visual results has to be adapted in a way that prevents occluding them with the user's finger or hand (see **goal 3**).

Local Navigation The second requirement of **goal 4** is local navigation. Previously discussed techniques, although being designed for global navigation, provide information that help with local placement of the lens as well. This includes term distributions as heat maps (Figure 4.3) or highlights when terms are moused over. However, as the adaption of local lens placement can be quite intricate, we include an additional local technique that supports users in confining lenses to the set of documents they are interested in. To provide information about the local quality of projection for the focused document sets, the stress measure could be used. As stress merely yields one single value for the projection quality of the area under the lens, the value for the entire 2D projection should be provided for comparison in such a scenario. Stahnke et al. [2016] provide an alternative by displaying joint errors for each projected point as halos of varying thickness that encode the amount of errors. With locally confined clustering, we pursue a different method

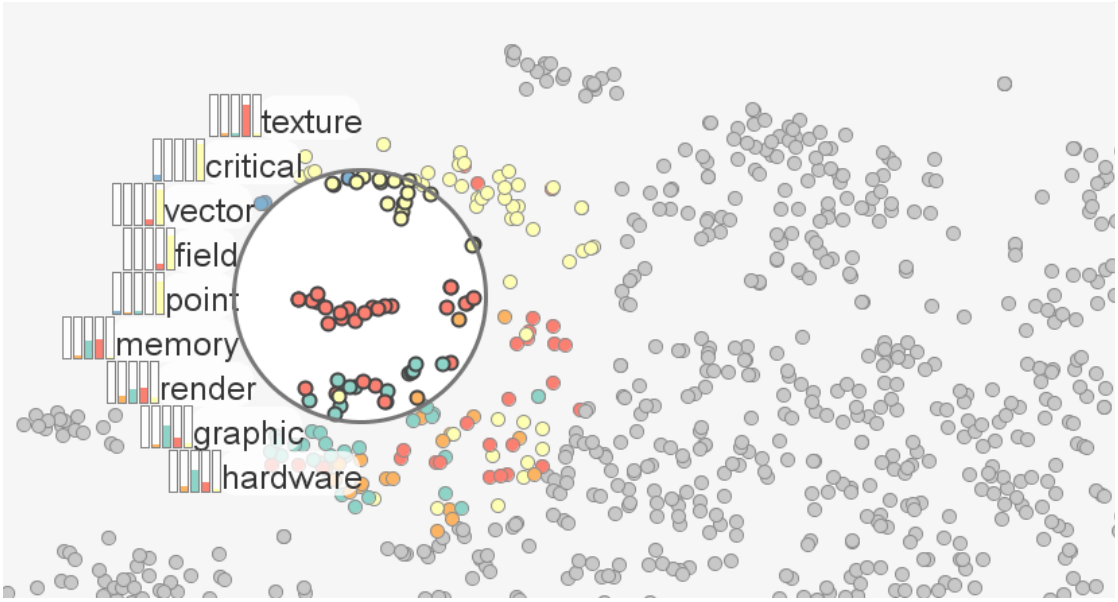


Figure 4.4 — A lens depicts clusters of documents by coloring document glyphs accordingly. Bar charts with respective colors are placed near the terms to indicate the relative prominence of the terms in different clusters. If users are interested in topics associated with *vector field* or *critical point*, they can move the lens up to include more related documents.

by grouping focused documents based on their high-dimensional representations. Once users activate this functionality, clusters are displayed by coloring document glyphs (see Figure 4.4). In addition, bar charts colored according to the clusters are shown next to each term. They indicate the relative prominence of the term in each of the clusters. This provides information about the similarity structure of the focused documents and helps with navigating locally. Learning about which of the documents are particularly close, and which are less close in the original space, users are supported in confining the lens to a smaller area that contains a higher rate of documents relevant to them. Feedback about the relevance of displayed terms for the clusters helps them to interpret the clusters and gauge their importance. In addition, as depicted in Figure 4.4, we extend the clusters to a small area around the lens. This can provide users with information about whether their focus is to narrow, and increasing the lens would include more potentially relevant items. Users are thus supported with optimizing the placement of the lens once they have discovered a region of interest to them. In addition, clustering provides information about the quality of the projection in the focused area. Many clusters scattered over a wide area of the layout might indicate a locally higher information loss through projection compared to many dense, locally confined clusters. Users

can adjust their exploration strategy to a more fine-grained approach in the former case, and a more coarse grained approach in the latter.

4.3.2 Document Characterization

Many methods have been proposed in NLP to summarize documents [Nenkova and McKeown, 2012]. Most of them distill longer texts into few representative sentences that summarize their content. We consider these methods largely unsuitable for text exploration with DocuCompass. The two main reasons are that we need characterizations of documents that are as brief and informative as possible, and that can be computed at interactive speed (see **goal 2**). Thus, we discuss term selection strategies that help users grasp the gist of a document set. We call these techniques *document characterization* in the context of this work to avoid confusion. Effectively supporting users with exploring, analyzing, and navigating through a large number of text documents is challenging. For users, it is important to get an overview of the main contents of a document set. This helps them to extract and extend information and knowledge about the corpus. It may also help to find additional entry points into the collection for further analysis. We concentrate on characterizing sets of documents based on a selection of terms they contain, selected according to different measures. Alternatively, users can activate various meta data lenses that show distributions of meta data attributes of the focused documents. In addition to being computationally feasible, term-based characterizations have the advantage of being quickly read and interpreted by users. Moreover, different types of texts, such as scientific literature, narrative texts, or micro blog messages are to be supported according to **goal 5**. The subsequently discussed techniques are highly flexible and can be applied to a wide variety of text types.

Term Rating The simplest term rating strategy is to use the document frequency (df) of the terms. As it emphasizes frequently recurring terms, df is suitable for large sets of heterogeneous short documents, such as micro blog messages. It does not work well for longer, uniform texts, such as paper abstracts from one particular journal or conference. For abstracts from the VIS community, e.g., terms such as *visualization* and *data* are selected for every focus set, as they appear in almost every abstract. In theory, this problem could be alleviated by creating domain specific stop word lists. A more practical alternative is tf-idf (see Chapter 2). It effectively removes common terms that have little discriminating information from the list. Another option is the G^2 measure (see Chapter 2). Especially for uniform data sets with many similar documents, G^2 helps to select those terms that help to learn about the idiosyncrasies of documents under the lens. It is thus a very helpful measure for content-based 2D projections, while it might not be the right choice for geo-located documents. Our DocuCompass implementation supports the

tf-idf and G^2 term rating method and allows users to freely choose between the two.

Linguistic-based Methods So far, we have discussed methods that rate and select terms based on their frequency of occurrence. In addition, methods based on linguistic knowledge about terms and their interplay within texts have the potential to increase selection accuracy. The most simple strategy is a stop word lists that comprises terms that do not contain information in isolation, and are thus not used for characterizations. This eliminates highly frequent terms, such as pronouns, conjunctions, and prepositions. Stemming and lemmatization [Manning et al., 2008] are another popular method to reduce the number of extracted terms. The former removes affixes from words and the latter reduces each token to a lemma, i.e., its base or dictionary form. Both methods conflate different morphological forms of words, thus reducing word vector dimensionality. Although lemmatization is a more expensive process, in contrast to stemming, it can handle irregular forms, such as "*went* → *go*". To display terms that characterize document sets, we have experienced problems with simple stemming. While it is mostly able to correctly identify all morphological forms of a word, it often reduces them to an ungrammatical form that can be hard to interpret. Displaying such terms can confuse users more than they help to understand the document contents. We thus prefer using full-fledged lemmatization in order to display grammatically correct forms and thus have included it into our DocuCompass implementation. Other methods to reduce and filter terms that have proven effective in Chapter 3, such as part-of-speech tagging (POS) or named entity recognition (NER) might also prove useful in this context. For example, for literary texts, frequently occurring characters might be useful characterizations, while for newspaper articles, mentioned locations might provide users with helpful information. Further methods exist that might be useful for specific types of documents or analysis scenarios, such as extracting and filtering technical vocabulary [Judea et al., 2014] for patent documents. None of this, however, is currently implemented in our prototype.

Metadata Another way of characterizing focused documents is through metadata. The types available are dependent on the documents. For scientific literature, metadata includes authors, their affiliations, the conference or the journal of a publication, the citations, and its index terms. Exploration of narrative texts might be enhanced with additional information of the characters, such as age, origin, relation to other characters, alternative names etc. When working with micro blog data, hash tags, linked persons, or the number of re-tweets are potentially important information. Metadata distributions in the focused documents can further provide important information for exploration. These distributions can be

displayed as plots or other types of visualizations next to the lens. For example, bar charts can depict the distribution of publication years for scientific articles (see Figure 4.1 (d)), or the number of citations over time.

4.3.3 Document Spatialization

The effectiveness of some of the characterization and navigation methods discussed for DocuCompass depend on the underlying document spatialization. For this reason, we list different spatialization types and discuss suitable lens configurations.

Inherent 2D Coordinates Placing documents into 2D is most straightforward if they are geo-located. This can be the case, e.g., for micro blog posts that contain the location they have been sent from. Other examples are patents, that can be placed at the location of their applicant, or hotel reviews, which can be placed at the respective hotel location. Using these locations directly results in scatter plots with well-defined axes that can be translated into latitude and longitude based on the map projection used. With geo-locations, one cannot generally assume that texts placed near to each other share any similarities with respect to their content. This calls for specific characterization techniques that are able to extract and convey the structure of a heterogeneous document set, including clustering and topic modeling. One way of doing this is to create geo-temporal clusters that help with the detection of outlier terms [Thom et al., 2012b]. These terms are extracted and displayed on the map and help users with global navigation. Terms that occur with unusual high frequency at a certain location may indicate a specific event that might be worth exploring. For effective navigation, methods that take the underlying map structure into account can be used. Depending on the user's interest, DocuCompass could support geographic navigation to, e.g., select all documents situated within the borders of a country, or a city. In other scenarios, users may want to move a lens along a road to analyze all micro blog posts sent by motorists. This can be achieved by lenses that snap to map features, such as political or geographic boarders.

Metadata-based Mapping Documents often do not have geo-coordinates or any other inherent spatial structure. For them, 2D mappings can be created from either metadata or textual content. In the former case, the simplest way to map documents into 2D is to select two metadata attributes that can be ordered in any way. The documents can then be laid out along those axis that have a clearly defined meaning. Examples of document scatter plots are, e.g., patents that are laid out according to the year they have been published on one axis, and the number of

citations they contain on the second axis. Micro blog messages, e.g., could be laid out according to their length, and the number of re-tweets. In such scatter plots, DocuCompass should be aware of the axis data type to allow navigation based on them. For example, the lens can snap to certain values or ranges of values, such as a time span on a temporal axis.

Dimensionality Reduction Another way to lay out documents is to optimize their pairwise distances in 2D to retain high-dimensional distances as well as possible. This typically introduces an information loss, as the distances cannot generally be mapped accurately into 2D. There is a number of different methods for such a mapping that vary in terms of optimization criteria and computational complexity. Some of them operate on high-dimensional vectors, while others take similarity matrices as input. Two methods that are currently popular due to their relatively small errors are LSP [Paulovich et al., 2008] and tSNE [van der Maaten and Hinton, 2008]. They typically operate on documents in vector space (see Section 2.2.2). Recently, methods to automatically learn term similarities have been proposed [Mikolov et al., 2013] and shown to achieve high document classification accuracy [Le and Mikolov, 2014]. This suggests that they might be useful for creating high-accuracy document spatializations as well. When using dimensionality reduction, the resulting axes of the 2D plots do not have any clear meaning. Thus, the first task of users during an explorative analysis is to understand a layout, and to find parts of the space that are of interest for deeper analysis. For this, DocuCompass is a paramount tool that allows users to change the size of the focused space and get a characterization of the documents according to many different criteria.

4.4 Implementation

We have implemented a software prototype that comprises many of the previously discussed methods. It is based on Java 1.8 and the prefuse library [Heer et al., 2005]. All NLP, including tokenization, sentence splitting, and lemmatization is done by Stanford CoreNLP [Manning et al., 2014]. In addition, the prototype includes content-based spatialization methods for documents, namely LSP [Paulovich et al., 2008] or t-SNE [van der Maaten and Hinton, 2008], based on their respective libraries. The prototype can be pointed to a text data set to read in, that may include the spatialization as metadata. Once the document scatter plot is created, users can explore it using the lens. Different lens types and characterizations can be activated through a context menu. The lens can be moved with the mouse, and its size modified with the mouse wheel. To keep DocuCompass responsive and ensure scalability even with very large corpora, we have integrated a quadtree data

structure that quickly retrieves documents in the focused region. The prototype currently offers tf-idf and G^2 for keyword selection that can be combined with additional cues that provide information about the focused documents. These include a bar chart that shows publication dates of documents over time (see Figure 4.1 (d)), if available in the data set. In addition, heat map previews are available for each of the keyword extraction methods. Hovering and selecting terms works as previously described in Section 4.3.1. Finally, for local navigation, the previously described clustering method has been implemented (see Figure 4.4). The cluster algorithm and its modification has been contributed to this project by Qi Han. It is based on the algorithm proposed by Rodriguez and Laio [2014] and runs at interactive rates, and with our own extension, estimates an optimal number of clusters for a given set of documents. The algorithm is based on identifying density peaks that become cluster centroids within the high-dimensional vector space, and subsequently assign all other documents to one of the peaks based on the density structure of the data. This has the advantage of producing relatively stable results when moving or resizing the lens. We estimate the optimal number of clusters by identifying good cluster centroids in high-dimensional space within the set of focused documents. Good candidates are selected based on high local density, and large distances to other centroid candidates based on the median absolute deviation method [Leys et al., 2013]. Automatically estimating the number of clusters frees users from the burden of having to provide any parameters for clustering during exploration.

4.5 Usage Scenario

In this section, we discuss a fictional usage scenario of our approach to provide a more lucid picture of it. The use case is based on a data set of scientific literature and sketches how the technique can be applied to scientific literature exploration. We base the example on our prototypical implementation of the technique, which currently offers a basic set of characterization approaches. Unfortunately, this limits the abilities of the approach somewhat, especially for such meta data rich documents as scientific articles. We still think that the usage scenario provides a more vivid account of the DocuCompass technique than mere technical description. This example uses the VIS publication data set [Isenberg et al., 2016], that contains all paper abstracts from the VIS(Week) conferences. At the time we were using it, the newest data was from the 2015 version of the conference.

An NLP researcher that is new to the field of visualization wants to learn more about the field. She has some basic knowledge about visualization, and now wants to delve deeper into the community and its different topics. She is particularly interested in recent research questions. She has also heard that there has been

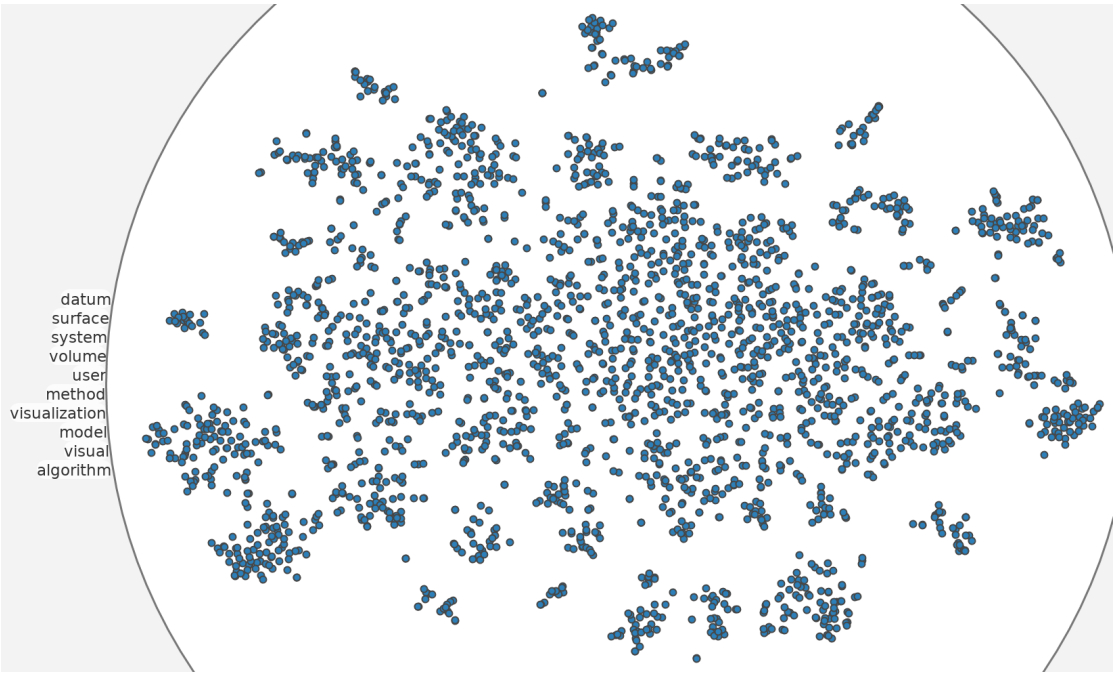


Figure 4.5 — A lens that hovers over the entire document set help the user get a general idea of the data set.

considerable work on visualizing uncertainties of statistical models recently, which she wants to know more about. After loading the data into the system, she starts to avidly explore it. To gain an overview of the data set, she first activates a tf-idf radial lens to explore the spatialization. She quickly recognizes visualization related keywords and terms, such as *rendering*, *treemap*, or *interface*. She enlarges the lens and moves it on top of all documents at once. Thus, she attains an overview of the most common terms in the entire data set, as shown in Figure 4.5. This brings up some generic terms in this context, such as *visualization*, *datum*, and *algorithm*, but she also discovers more concrete topics such as *volume* and *surface*. To look deeper into the different topics in the data, she activates several G^2 lenses to inspect the distinguishing features between the focused visual cluster of documents and the rest of the data set. She explores the data set at different granularity levels to analyze multiple visual clusters and areas of the space. Each time she identifies a topic that draws her interest, she uses the navigation aids to close in on it, and then marks it by leaving the lens at the respective spot. Continuing with this procedure for some time, she gradually marks several spots that comprise documents that, as judged by the terms yielded for them, belong to different research areas in the field of visualization. The final state is depicted in Figure 4.6, with the identified clusters highlighting documents from the following areas: volume rendering (Figure 4.6

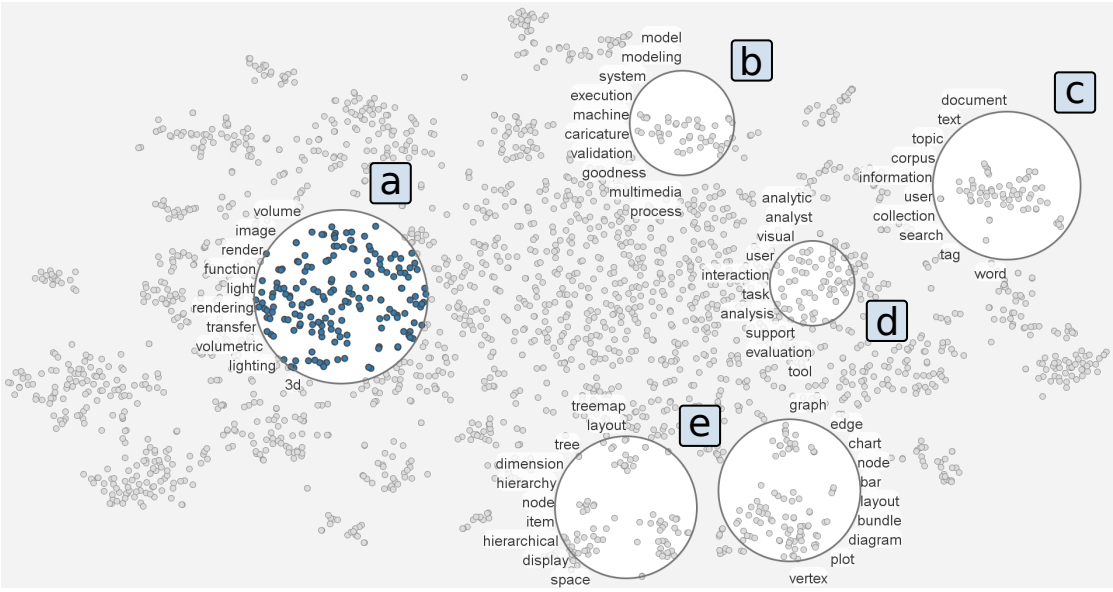


Figure 4.6 — Multiple G^2 lenses display the main terms of different clusters in the spatialization, pertaining to different research topics in visualization. The documents of the currently activated lens (a) are highlighted by color. (a) Volume rendering, (b) fitting and validation of models, (c) text visualization, (d) interactive systems to support decision making, and (e) tree and graph visualization.

(a)), fitting and validation of models (Figure 4.6 (b)), text visualization (Figure 4.6 (c)), interactive systems to support decision making (Figure 4.6 (d)), and tree and graph visualizations (Figure 4.6 (e)).

Another cluster she encounters seems to contain articles with research in the area of uncertainty visualization. She wonders about the recency of this research. To learn more about it, she activates a lens with a bar chart that depicts the distribution of publication years of the focused documents. She finds out that, particularly in recent years, uncertainty visualizations seem to have gained popularity (see Figure 4.7). Now that she has identified an area that contains texts that are particularly interesting to her, she wants to know if there is more potentially interesting material in the data set. To get a first glance, she activates the mini heat maps, and finds that there are additional areas with few occurrences of the term. She learns more about this after selecting the term *uncertainty*, which highlights all documents that contain this term in the entire spatialization (see Figure 4.8). The user looks into these examples and explores their local context in more detail. She learns that *uncertainty* appears in different contexts and topic areas in the data set, and play a role in, among others, the area of visual interactive

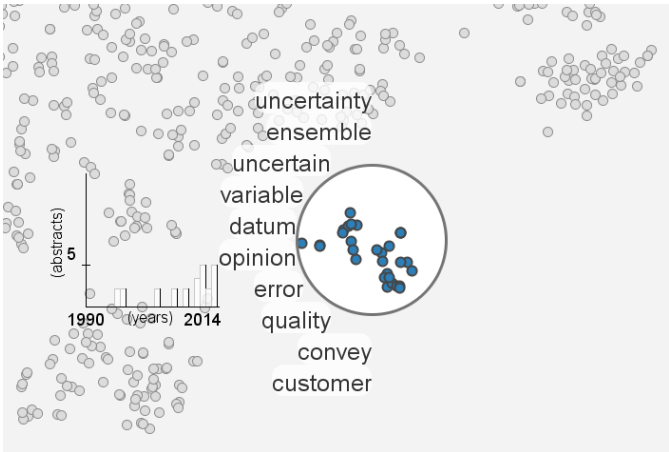


Figure 4.7 — A lens with a bar chart to depict the distribution of publication years.

machine learning, visualization of scientific simulation data, and visualization of biological data. However, despite these examples, not many articles outside of a small cluster seem to deal with uncertainty information. In fact, there is none in the area of text and document visualization, whose position she remembers. Nevertheless, she is satisfied with her results, and decides to take the documents from the *uncertainty* cluster she identified under closer scrutiny to learn more about specific uncertainty visualizations. Of course, as an NLP researcher the text and document visualization cluster from Figure 4.6 (c) has also aroused her interest, and she further plans to look into these documents, too.

4.6 User Feedback and Discussion

Generating insights into our interaction technique through user feedback is difficult. Beyond the typical problems that make visualization and interaction evaluation challenging [Plaisant, 2004], many facets influence the effectiveness of DocuCompass. These include the choice of text processing, the visual representation of the results, the type of text used, the spatialization, etc. A thorough comparative user study requires a large number of test sessions, and different corpora would have to be used to rule out learning effects. Testing explorative interaction techniques designed without clear information needs is particularly challenging, as there is no definition of what successful exploration is. Being able to describe all topics in the document set? Being able to identify some of them? Results can vary greatly depending on a test subject's previous knowledge making comparisons between subjects difficult as well. Even characterizing insights [North, 2006] is not applicable to such open-

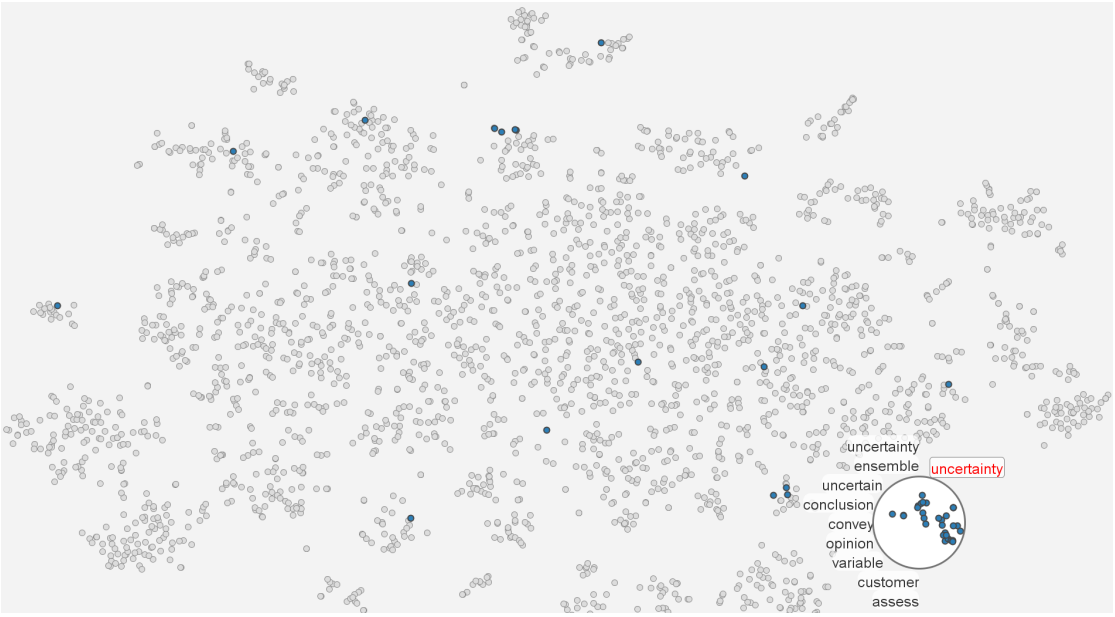


Figure 4.8 — A lens with the term *uncertainty* selected, and the documents that contain this term highlighted.

ended exploration tasks, as the insights and their complex interplay in deriving knowledge cannot be measured accurately. As a consequence, we decided to do a think aloud study with different user groups in order to collect and reflect on their feedback. Despite these limitations, we find that solicited feedback and the advantages discussed below indicate the effectiveness of the DocuCompass approach.

4.6.1 Software Prototypes

We have designed the think aloud study as a comparison between the DocuCompass technique, and an approach based on inspecting single documents. While the DocuCompass allows users to adjust the granularity of their exploration of the data set, inspecting and characterizing single documents constrains them to just one fixed level. For this, we created two software prototypes that could be used by participants. Both displayed exactly the same spatializations of the three corpora we used for the study sessions, but offered different exploration methods. As we aimed at soliciting insights about the lens technique, we decided to keep the spatialization fixed during user explorations. We used tSNE [van der Maaten and Hinton, 2008] to project all of the data sets that we presented to the participants as it is a state-of-the-art technique for dimensionality reduction. The first prototype

supports the selection of single documents from the scatter plot by hovering over their respective glyph. Selecting a second document automatically releases the initial selection, thus users can only focus one document at a time. Next to the spatialization, on the right side of the screen, two text areas show information about the focused document. While one of them contains its full text, including document titles, the other one characterizes the current document with a word cloud. Users can choose between tf-idf and G^2 for term selection. In addition, hovering over a term in the word cloud highlights the documents that contain it in the spatialization. The second prototype implements a basic version of DocuCompass, with reduced functionality. We mainly deactivated local navigation support to allow for a fair comparison with the previous prototype which does not offer any local navigation features either. In the second prototype, users are able to activate several different lenses, that, based on tf-idf or G^2, show a selection of the ten most highly ranked terms from the focused documents. In addition, term distribution previews can be activated next to the terms. For both of the prototypes we have refrained from including static labeling of the 2D document space. The reason for this is that it would have been difficult to attribute findings to either the static labels or the DocuCompass technique.

4.6.2 Participants and Procedure

Overall, nine persons participated in the sessions, one female and eight males. Their average age was 31 years (between 27 and 39). Three of them were computer science PhD candidates, two from the field of visualization, one from the field of computer vision. While the former two had a strong background in information visualization and knew the magic lens technique well, the latter had never heard of this concept. Three other participants were in the field of NLP, two as post-doctoral researchers, and one as a PhD candidate. All of them had basic knowledge of information visualization and one of them had heard about magic lens technique. The remaining three were M.Sc. students of mechatronics and food engineering with no background in information visualization, and none of them had heard about magic lenses. Individual study sessions with each participant lasted for about 30 minutes, depending on completion times and the length of the subsequent discussions.

We prepared three different data sets for the study. The first one was used for an introductory session and contained all paper abstracts of VIS publications [Isenberg et al., 2016]. The second and third data set were Reuters news wire texts selected according to their topic (sports news, and international conflicts) from the Reuter's RCV1 corpus [Lewis et al., 2004]. We permuted each of the four possible combinations of implementation and data set, and asked each participant to complete an

exploration task with both implementations on different data sets. The ordering of the techniques was counterbalanced in the user sessions. We posed the same three questions, independent of the data set or prototype used, at the beginning of each session. The questions were chosen so that users freely explored the data set during the sessions and focused on different levels of granularity. They asked about the general theme of the entire data set, the topics that users had identified within the spatialization, and for any subtopics into which these topics could be split up. These questions were to be answered by the participants once they had finished the exploration.

The exact procedure for each study session was as follows: **1)** We asked the participant to fill in a form with information about their person, their professional background, and their experience with magic lenses. **2)** Each participant received an introduction to both systems with the VIS abstract corpus loaded. **3)** The participants were asked to complete the first session. **4)** Then, they were asked to complete the second session on a different implementation and data set. **5)** We solicited oral feedback about each of the implementations and their features using Likert scale questionnaires, and led discussions about possible areas of application, and extensions to DocuCompass. We conducted the study session according to the think aloud paradigm, asking participants to voice any of their thoughts during the sessions and recording everything on paper.

4.6.3 Results

In the following, we report and comment on the results of the user feedback sessions. All of the participants were able to successfully use both implementations to solve the exploration tasks. All of them perceived the lens approach as faster and more effective. As this was part of our design goals for DocuCompass, this result was expected. In addition, six of the nine users mentioned that they were surprised at the speed of the lens and the absence of any lags when moving it. One of the most popular strategies used by five of the nine users to solve the exploration task was to start with a large lens, hover over all of the documents at once to get a general idea of the data set, and then shrink the lens to focus single visual clusters. This nicely illustrates the effectiveness of lenses for seamlessly switching between granularity levels. Seven of the nine users positively mentioned the possibility of getting a quick overview with the lens, and its effectiveness for the analysis of single clusters. For the sports data set, two users mentioned that titles are often quite meaningful as they contain the type of sports the articles are about. Although we included titles within the text that the listed terms were extracted from, visible titles improved cluster exploration speed with the lens-free prototype. This effect,

however, was diminished by a general reluctance of participants to read complete texts.

As we expected, the NLP researchers all found that hovering over a term in the word cloud and view its distribution is very helpful in the simple system, but they rated the lens approach overall as faster and more pleasing. Furthermore, DocuCompass was rated as being very scalable and versatile in that it can be applied to a wide range of different text types. In addition, one person noted that switching between views was quite arduous in the simple approach. Two participants praised the lens for being an excellent tool to explore visual clusters in 2D layouts. The mini heat maps were not used, as the NLP specialists generally thought that hovering a term yields a more lucid visualization of its distribution. While this is certainly true, the result came as a surprise to us, as they did not even use the heat maps to get a first peak at the distributions. In addition, only one of the NLP experts found using multiple lenses at once helpful. This surprising result might indicate that using lenses effectively requires some experience. Comparing these results to the visualization researchers, who generally liked using multiple lenses corroborates this. All of the experts in this group can think of use cases for these types of lens systems, two of them would even use it for their own research.

The computer science PhD students all found the lens to be helpful and effective for exploration. As expected, the term list was rated as being very useful for gauging the themes of documents underneath the lens, even though two persons lamented missing direct links to the documents. We did not expect this remark, as it would lead to significant clutter, considering that we only show the top ten terms. The same experts rated the different levels at which the lens is able to explore fine-grained as well as coarse topics as one of its great advantages. Two experts found the possibility of using multiple lenses at once as being very beneficial for analysis. Further, one visualization researcher made a positive remark about the interactive and fast reaction times of the prototype. Two participants mentioned that the small heat maps and the possibility of mousing over terms is a vital function for effective exploration, as it helps to get a fast overview of terms across the layout. The third expert found that the small heat maps take some time to process mentally and to map them to the large map. He thus rated them as little helpful. This remark did not come as a surprise to us, as we made similar experiences ourselves. However, we found the mini heat maps useful for quickly reviewing and comparing term distributions. The remarks from the other two participants insinuated that they agree with this. All three positively mentioned the stability of the terms and their ordering when moving the lens, in opposite to the per-document word clouds of the lens-free system. They could all think of tasks in which text lenses would help them with their daily work.

The feedback from the B.Sc. and M.Sc. students was largely in line with our expectations. They found the lens to be an easy-to-use, versatile, and visually appealing tool for exploration of texts that provides quick overviews of data sets. They could all think of applications for many different types of texts. To get an idea of how regions of the space are related to the currently focused documents, they found the mini heat maps particularly helpful. In addition, mousing over terms helped them to find new regions for analysis. One of them remarked, that it would be helpful to view more than the ten most highly ranked terms in an extra view to avoid occlusion problems. Of course, in addition to all feedback pertaining to the approach itself, some minor usability issues were noted, and possible extensions suggested. Those were, for example, about the color mapping we used in the sessions, that terms can become too long and occlude much of the space next to the lens, and that activating lenses in the prototype is a bit complicated.

4.7 Discussion

The results indicate that the proposed lens techniques are effective for the exploration of document spatializations. Even with the same term extraction method, the DocuCompass approach shows a more stable term list compared to the word clouds for individual documents. This is because DocuCompass aggregates over several documents and depicts their similarities. The effect obviously helps users to grasp the topics or themes more quickly. We attribute the positive emphasis of the fluid interaction experience by our study participants to two factors: the usage of a quadtree data structure for speeding up the computation of characterizing terms and the design choice to depict a suitable number of labels. The flexibility of the lens to analyze document sets of different size supports users in exploring spatializations. This can be derived from the exploration strategy adopted by several users. Approaches that only offer a predefined set of abstraction levels might decrease exploration effectiveness. The users had mixed opinions about the navigation aids we included. While the possibility of mousing over terms and getting an overview of their distributions was rated high, the mini heat maps were only used by few users. One possible reason for this could be that the mini heat maps take some time to mentally process and map them to the large map, as one participant mentioned. The study also provides insights for future work. One participant mentioned that it would be helpful to be able to do on-the-fly switches between different visual characterizations with DocuCompass during analysis, instead of having to use a different lens. This is certainly true and we plan to implement this in the future, since this supports a more flexible analysis. Another remark was that multiple lenses might be more helpful if relations between them were shown.

Several participants proposed concrete application scenarios, e.g., from the digital humanities, where DocuCompass could be applied.

We have developed DocuCompass out of the need for an interaction technique that offers insights on intermediate levels of granularity when exploring document spatializations. Very coarse labeling on the overview level as well as very detailed information on a per document level was simply too shallow or too detailed for many tasks we encountered. A focus+context technique is one possible solution to solve this problem in a scalable way, but also poses challenges. Scalability is an important factor in many respects.

Interaction scalability Focus+context interaction has to be fluid. Otherwise, working with large document collection becomes cumbersome and tedious. Depending on corpus size and the applied characterization technique, on-the-fly processing of documents can be slow. There are several ways to speed up interaction. Hierarchical data structures that speed up access to the required information can be used, such as quadtrees, kd-trees, and others, to efficiently access focused documents in spatializations. In addition, results for text processing can be pre-computed and integrated with these data structures by aggregating information at intermediate nodes to speed up expensive mining tasks. This shifts computational effort to the pre-processing phase, which is often worthwhile for static document sets.

Information scalability The ability to extract relevant information from large data collections is important to facilitate exploration tasks. Depending on the task, different modes of aggregation are appropriate to achieve this. Users with murky information need can follow different strategies during exploration, depending on their new insights and aspects of the data set they develop an interest for. Possible strategies include general exploration of themes, comparative analyses of different parts of the set, berry picking strategies, and specific aspects such as viewing the persons that appear within documents. To support multiple exploration modes, DocuCompass offers different text processing techniques. Whether it is suitable to let users decide which technique to use or to restrict it to a specific one, depends on the task, scenario, and their expertise. One aspect we underestimated when developing first variants of the technique is the difference between text types. As discussed previously, they significantly influence the usefulness of certain mining techniques.

Visual scalability We use text labels as cues, since they are a natural choice to represent text documents. The associations and background knowledge that humans link with words make them much more powerful than other visual cues. The interpretability of the focused documents increases with the number of terms shown.

This makes it possible to disambiguate complex content and develop an idea of the underlying document set. Even with no further context of the terms within the document collection under the lens, they seem to more effectively convey information than quantitative values of structured information. However, the uncertainty introduced by neglecting contexts poses the risk of misinterpretation. Different term extraction techniques might counteract this issue. And by changing lens size and position, additional information can be quickly explored to correct inaccurate interpretations. The quality and interpretability of characterizations using labels also depends on the length of texts, their number, type, and spatial distribution. Monothematic texts and content-based spatializations can be characterized more easily and coherently.

We decided to depict only a fraction of terms to make interpretation simple and exploration fast. The omission of information and the corresponding risk of misinterpretation could be indicated with additional cues that show the severity of omissions. This can convey a notion of uncertainty regarding the analysis of the currently focused documents to the users. We consider such techniques part of our future research. How quickly text labels can be perceived and interpreted by humans as opposed to other visual representations is a different question. In DocuCompass, the labels remain relatively stable if the lens is moved, but this, of course, depends on movement speed and the distribution of documents. When using DocuCompass, we found that labels of equal size are perceived faster than those that encode additional aspects with font size. This led to the decision to represent prominence or importance by the order of the labels. We assume that keeping the position of the labels stable in relation to the lens expedites their interpretation. Still, further assessment of the shown number and size of labels is required to optimize the visual cues for perception.

4.8 Future Directions

With DocuCompass we have developed a straightforward exploration method for document spatializations. Its lens-based design has the advantages to support continuous exploration tasks. It thus fills a gap between visualization and interaction techniques that provide large scale overview or detailed inspection of text corpora. DocuCompass can also be easily extended with almost any text analysis procedure. It is thus flexible and adaptable to different text types and exploration and analysis tasks. In addition to a round shape, future implementations could offer user defined shapes, when an adaptation to complex document distributions in the spatialization is useful. In this chapter, we discussed a limited set of visual cues to characterize document sets. While we consider text labels as the most informative ones, many other possibilities are thinkable, and we have demonstrated some of them previously.

During some stages of exploration, e.g., when users actively search for specific information to extend or test a frame during sensemaking (see Chapter 2), user analysis gets more targeted. This is supported with different navigation aids. We did not discuss the transition from exploration to a more targeted analysis. A straightforward way to realize this can be the depiction of focused documents in a separate view. To explore and analyze them further, additional approaches targeted at such an analysis could be included, such as the ones from the previous Chapter 3. Making this transition as smooth and effective as possible will be part of future research endeavors.

CHAPTER 5

Interactive Classifier Creation and Exploration

This chapter contains previously published material from the following publications:

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Transactions on Visualization and Computer Graphics*, 18 (12):2839–2848, 2012b. doi: 10.1109/TVCG.2012.277

S. Koch, F. Heimerl, and T. Ertl. Visual document retrieval: Supporting text search and analysis with visual analytics. *Computing in Science & Engineering*, 15(4):66–74, 2013. doi: 10.1109/MCSE.2013.93

H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *Transactions on Visualization and Computer Graphics*, 19(12):2022–2031, 2013. doi: 10.1109/TVCG.2013.186

Until now, we have been looking at two approaches that facilitate the exploration of text and document data sets. These approaches support free exploration and

corresponding human sensemaking processes that help users attain knowledge about the properties and contents of text data sets. However, both approaches do not support storing results or insights created from exploring text data. This chapter is an excursion into scenarios in which machine learning approaches and human users learn from each other during exploration. This enables a user not only to retain gained knowledge for later analysis, but also to transfer it to a machine learning method through interaction during the exploration process. The concrete example we are looking at in this chapter is the interactive creation of binary text document classifiers. This is not a pure exploration approach, but it strongly supports explorative analysis of the machine learning results and the base data set.

5.1 Motivation

Search and retrieval tasks can play an important part during exploratory analysis. This is especially true for the analysis of textual data, for which human interpretation and world knowledge is crucial. Filtering and identifying relevant data items that directly cater to the user's current cognitive sensemaking processes (see Chapter 2) can expedite knowledge extraction from textual data by supporting foraging processes [Pirolli and Card, 2005]. After initial exploration of a data set, users often reach a point at which they develop a particular interest in a certain subset of the data set. Depending on the VA approach used for exploration, the set of desired documents is not necessarily accessible in a straightforward way. With the approach presented in Chapter 4, for example, the available 2D mapping may or may not group documents relevant to the user at a specific point of exploration into one common area. Let's assume that the user develops an interest for publications that include a full user study while exploring a set of VA articles. She then, of course, wants to explore the subset of documents that contain such studies to compare and review them in more depth. Automatically identifying those documents that exhibit aspects that the user is currently interested in helps her to quickly shift her attention to this exact subset, which is conducive to multiple forms of sensemaking (see Chapter 2).

One way of supporting users with identifying all documents relevant to such an information need are traditional text retrieval methods [Manning et al., 2008]. These are represented by the well-known search engines on the web such as Google, Bing, and Yahoo and based on text queries formulated by users. Such methods, however, are largely unsuitable for exploration and analysis scenarios that we aim to support. There are two main reasons for this. Firstly, while typical document retrieval scenarios assume one piece of information that the user needs to satisfy her clean-cut information need, retrieval scenarios in analysis contexts are typically

recall oriented [Pirolli and Card, 2005]. This means that all documents relevant to a specific topic or aspect have to be collected to ensure a comprehensive analysis. In such scenarios, users tend to trade a lower precision of the search method with a very high recall lest they miss something important from the underlying data set. Secondly, developing an interest into one specific aspect or topic of the data set during exploration does not yield crisp and well-defined information needs. Often, the search criteria of what kind of texts the user is looking for manifests during exploration, and then solidifies and sharpens at the same time. This calls for an interactive retrieval method that supports explorative analysis and query refinement at the same time.

For this reason, we propose the use of statistical text classification for retrieval in such scenarios. More precisely, we present an interactive visual approach that supports the exploration of text classifiers that find documents according to criteria defined by the users through example texts. The approach is based on SVMs (see Chapter 2 for details) as the base text classifier that is interactively created by the user through exploring the underlying text data set as well as the state of the classifier. This helps them to identify and label positive and negative examples of what types of document are relevant to them, which then, in turn, helps to improve the SVM classifier. The ensuing iterative process that switches between exploration of the classifier enables both user and machine to learn from each other. While exploring new examples that the classifier digs up, users come across and learn about new aspects of the data set and their still murky information need. At the same time, human sensemaking and language understanding skills help them to see connections, or draw conclusions about the data set and its contents that provide them with new insights and knowledge. With the presented approach, they can instantly return this knowledge to the SVM algorithm in the form of positive and negative examples of what documents they consider relevant. In order to keep this interactive loop as efficient as possible, we base our approach on principals from active learning (AL; see Chapter 2).

Being able to quickly create an SVM classifier that can separate documents according to specific aspects is not only important to allow users to concentrate their attention on exploring subsets interesting to them, but can also serve as a means to explore dynamic data sets. The previously mentioned user interested in user studies, for example, can apply the classifier to newly published articles and papers and continue exploring them. Thus, new developments or dynamic changes within the data set can be successively uncovered and explored. In addition, exploring and supervising the SVM algorithm during its learning iterations helps users gauge the accuracy of its results, and ultimately increase trust in it. The visual approach that we present in this chapter starts with a keyword query that bootstraps the initial classifier. A user can thus formulate her partial information need as a keyword query

that yields relevant documents. We added this mechanism to create a stand-alone application of the approach that we could use for evaluation. However, instead of bootstrapping the classifier by using a query, one can easily think of different possibilities of including the presented approach into an exploration system such as the one from Chapter 4. Selecting a specific term, using the documents under the lens, or identifying and selecting a few relevant documents from different areas of the spatialization are all straightforward interactions to bootstrap a classifier.

5.2 Related Work

Multiple approaches have been proposed that support the exploration of classification algorithms or the interactive labeling of training sets for machine learning. Seifert et al. [2010] support users with exploring document collections to find good labeling candidates for classification. They employ a document map visualization method that has been generated through data clustering. Moehrmann and Heidemann [2012] present a systems that uses unsupervised learning to facilitate fast labeling of image data sets for classification. We think that combining supervised classification methods with unsupervised clustering or topic extraction may lead to two competing approaches. If there are automatic methods that can already identify clusters the way users need them, it is better to use those unsupervised methods, because there is no need to label training data. Only when users need their own, self-defined categories that are not reflected by any automatic clustering investing work into creating a classifier is a viable option. Similar to our approach, Seifert and Granitzer [2010] present a visual technique that reverses the initiative in the AL loop from the machine to the user. It displays unlabeled examples in a multi-classification scenario. There are three main differences to our work. Firstly, their visualization is solely based on the confidence of the classifiers' decisions. Their radial layout concentrates uncertain documents at the center which complicates interaction. In contrast, we relax the layout of documents along the decision boarder to help users with labeling instances that bear the highest potential to improve the classifier. We also provide additional information about the underlying document collection as well as the current classification model. Secondly, they use a standard seed set for their experiments, while we propose a bootstrapping approach. Thirdly, they evaluate their approach by simulating user behavior, while we conduct an evaluation with real users.

Various ways of directly incorporating a user into classifier training and model evaluation have been discussed. Ankerst et al. [2000] describe the goal of such a cooperation between users and computer as the full exploitation of the capabilities of both. Bertini and Lalanne [2009] term this cooperation between user and computer "interactive machine learning". Endert et al. [2011] coin the terms *parameter level*

interaction and *observation level interaction*. The former allows full visual steering of an algorithm's parameter, while the latter only supports interaction on the instances it operates on. Our approach is mostly based on the latter concept, but also has some features that provide insight into the algorithm's parameters. May and Kohlhammer [2008] describe a general approach to create customized classifiers in an interactive visual manner and sketch it as an extension of the information visualization pipeline. They exemplify the approach by enabling users to build and iteratively refine a classifier based on a decision tree. Interactive decision tree construction is a common variant of visual interactive machine learning approaches [Ankerst et al., 1999; Ware et al., 2002; Liu and Salvendy, 2007; van den Elzen and van Wijk, 2011]. The systems enable users to directly manipulate their models, e.g., by creating new tree nodes and selecting attributes and split points for them. One reason for the popularity of decisions trees in interactive approaches is certainly that their models can be easily transfered to a well-understood visual analogy. Höferlin et al. [2012] present a system tailored to the ad-hoc training of classifiers during video surveillance. Poulet [2008] presents multiple methods for visualizing classification models, including decision trees and SVMs, but only support limited interaction with the data. An approach to interactive regression analysis for rating network security events is presented by Eaton et al. [2009]. Mühlbacher and Piringer [2013] also support the interactive creation of regression models for arbitrary application scenarios. Fails and Olsen [2003] present an interactive approach to create classifiers for images. Fogarty et al. [2008] develop an image retrieval system that allows users to steer the creation of classifiers for different concepts, e.g., scenic image, or images with faces.

Our approach has similarities with relevance feedback [Ruthven and Lalmas, 2003] in information retrieval, because it allows users to provide feedback based on relevant and non-relevant examples. In contrast to these systems, our approach yields a classification model that can be stored and used after it has been trained once. Recently, there have been further approaches that incorporate interactive machine learning. Gleicher [2013] use an SVM algorithm to automatically explain user-defined categories of data instances based on its data attributes. Behrisch et al. [2014] employ machine learning to classify different scatter plot visualizations of a high-dimensional data set based on the user's relevance ratings. The system then yields scatter plots of variable correlations that are of potential interest to the user. Alsallakh et al. [2014] provide a visual exploration approach for the results of multi-classification that allows users to identify and solve problems with the performance of classifiers. Paiva et al. [2015] propose an interactive classifier creation approach that supports labeling of instances to incrementally develop classification models.

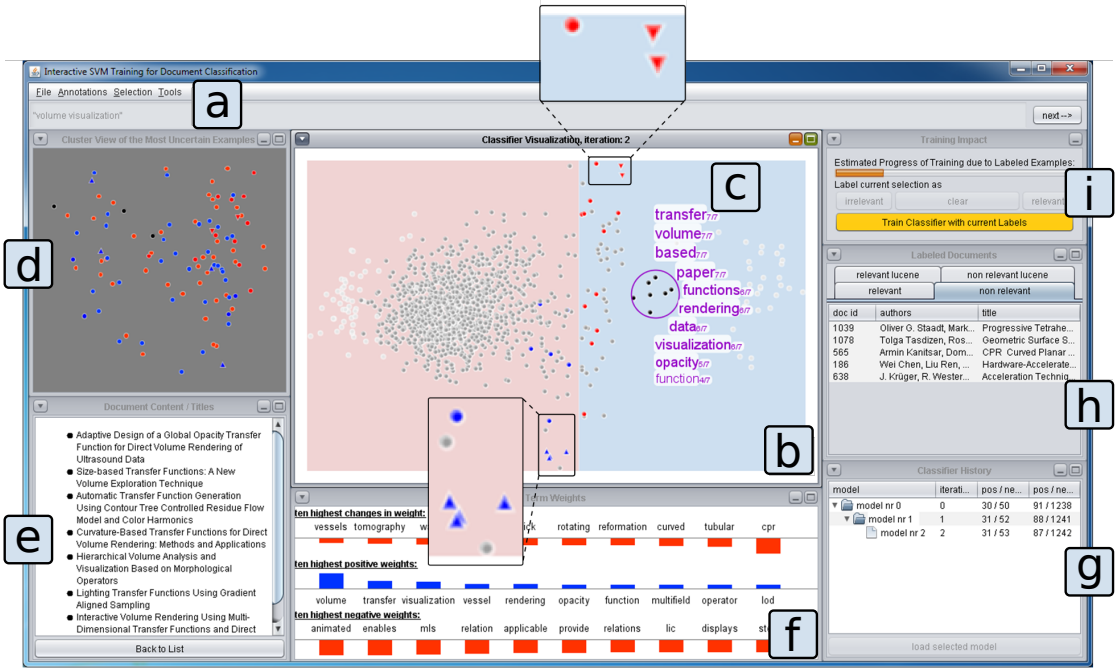


Figure 5.1 — Workbench to interactively create an SVM classifier during document exploration. It consists of (a) text field for bootstrap query, (b) visual representation of classifier state and the document set, (c) term lens for exploration, (d) cluster view to specifically explore uncertain documents, (e) content view for document texts, (f) term weights view of the classifier, (g) classifier history, (h) labeled documents view, and (i) controls for the labeling functions.

5.3 Interactive Classifier Creation

Figure 5.1 shows our classifier creation desktop. The process of creating a classifier in the current stand-alone implementation is depicted in Figure 5.2. Before the interactive process is started, users have to enter a keyword query into the text field (Figure 5.1 (a)). This query is executed on the entire document data set based on the Lucene[1] information retrieval library that stores documents based on tf-idf representations. The 50 most highly rated positive results are used as positive training instances, while 50 randomly selected negative results from the remaining data set are used as negative instances. Loading the initial training set from a file is also supported by the current implementation. This initial training set is then fed into the SVM learner, and the initial classifier is created from it. Once

[1] http://lucene.apache.org

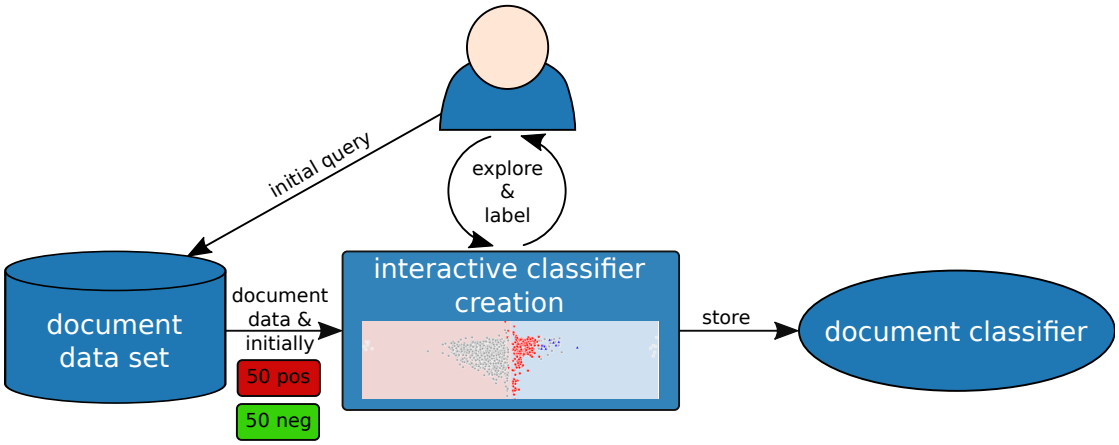


Figure 5.2 — Current stand-alone implementation of our interactive classifier creation approach. The user bootstraps the process by providing an initial keyword query. A retrieval mechanism then extracts the 50 most relevant, and 50 random non-relevant documents to bootstrap the classifier. During an interactive process, the user refines the classifier, and finally decides when to stop and store or apply the classification model.

the classifier is created, in is depicted within the classifier state and document set visualization (Figure 5.1 (b)) of the desktop. In it, the classifier's current state with respect to the entire document set is depicted, including training documents as well as unlabeled ones. The documents depicted as white points are examples that have already been labeled, and the gray ones are unlabeled. The view is a 2D representation of the high-dimensional classification space, with the two sides of the SVM's decision boundary being colored red, for the negative side, and blue for the positive side.

Users can explore the classifier's data interactively and gain insights into the data set and the documents it contains. Here, the exploration process of the data set is resumed, enabling users to deepen their knowledge of the data set, and continue their sensemaking processes about various aspects of the documents. For this, the view includes a text lens (Figure 5.1 (c)) for exploration that has been designed according to the paradigm presented in Chapter 4. Similar to this implementation, and for the same reasons, the lens shows the ten most prominent terms of the documents under it. In the current prototype, document frequency (*df*; see Chapter 2) is implemented as the prominence measure for the extracted terms, but other measures (see Chapter 4) could be implemented, too. In addition, to give users additional information about the prominence of terms in the focused documents, the *df* for each of the depicted terms are shown right next to it. Exploration of the state of the classifier, however, is also an important part here.

To support this, the approach is designed to comprehensively provide insight into the current classification model. For this, we create the layout of the 2D plane in accordance with the current classifier as follows. The horizontal axis of the plane represents the distance of the documents to the decision hyperplane in high-dimensional dictionary space. This allows users to accurately gauge the uncertainty of the classification. Uncertainty is lower for documents that are further away form the decision boundary. On the vertical axis, the documents are distributed according to similarity, based on the first principal component of the 100 most uncertain documents. The reason we do this is twofold. Firstly, determining the principal component is computationally rather expensive, and has to be done each time the classification model is updated, and the instances are laid out according to the new model. Secondly, in order to achieve a particularly high similarity resolution of the documents close to the hyperplane, which are the ones with the highest potential for classifier development, we have decided to determine their first principal component. For this small set, computation is possible at an interactive rate. The documents further away from the boundary are then laid out according to their similarity to these high-uncertainty examples as follows:

$$v(d_i) = \frac{\sum_{d \epsilon U_i} e(d_i, d)^{-1} \cdot v(d)}{\sum_{d \epsilon U_i} e(d_i, d)^{-1}}$$

where $v(d_i)$ is the vertical position of an instance d_i, U_i are the ten instances from the high-uncertainty set closest to d_i, and e gives the Euclidean distance in the original vector space. Emphasizing the most uncertain documents is an aspect of the approach that is based on traditional AL techniques. While learning about the data set and the classifier, users label relevant and irrelevant documents (red/blue triangles visible in the enlarged areas in Figure 5.1) according to their forming and solidifying information need. To better gauge the effect of their labeling actions, and get instantaneous feedback, the documents that would be classified differently (red/blue circles in Figure 5.1) if the classification model was retrained with the currently labeled documents, are also highlighted. Their color indicates their new class.

During exploration, the provided views and mechanisms guide users to label documents that promise considerable training progress and provide feedback about the estimated impact of the progress. This strategy is, again, based on AL (see Chapter 2) to speed up the creation of the classifier, and significantly reduce the number of labeled documents for a satisfactory classification performance. The cluster view (Figure 5.1 (d)) provides additional information about the most uncertain documents to help users get a more comprehensive account and to support sensemaking by identifying potentially inconsistent or erroneous classifications. It shows a 2D clustering of the 100 most uncertain instances based on the LSP

algorithm [Paulovich et al., 2008], without depicting the classifier's decision boundary. The layout thus solely reflects document similarity to support sensemaking, and shows classification decisions as colors of the document glyphs. Promising regions to look for labeling candidates are tight clusters of heterogeneously classified documents, as these are likely to contain mislabeled instances. Labeling such documents helps the classifier learn new properties that help to distinguish between both classes. For effective exploration, the term lens with the same configuration as described above is available in the cluster view as well. In this view, users can also select and label documents. Every interaction with the classifier view and the cluster view is reflected in the other, including selection and highlighting operations. The same instances in both views are thus easily identified by users.

The content view (Figure 5.1 (e)) displays an entire document in focus. If more than one document is focused, e.g., by using the lens, a list of their titles is displayed. By clicking on an individual title, users can view the contents of a single document. This allows users to drill down to the actual documents, for example, if it is hard to decide for a particular labeling candidate whether it is relevant to the user's information need. User can also get deeper insight into the classification model with the term weights view (Figure 5.1 (f)). As the underlying classification method is a linear SVM, the model computes weights for each of the terms that appear within the document set. These weights can be positive or negative, depending on what class they are indicative of. To represent these bases for classifier decisions, we show the ten highest positive, negative, and, based on the difference to the previous SVM model, the ten highest changes in weights. All weights are depicted as bar charts, so that the relative differences can be compared easily. These bars can also be hovered or selected, which then shows or selects all documents that contain the respective term in the classifier view. This provides feedback about the distribution of these features within the data set. The classifier history (Figure 5.1 (g)) supports sensemaking by allowing users to track multiple lines of analysis, and return to previous states at any time. It contains a tree of all previously created and explored models and provides some basic statistics for each of them as an overview. To have direct access to the current training set, the labeled documents view (Figure 5.1 (h)) lists all instances that have been labeled. The document lists are divided according to the labels, and whether the documents have been labeled by the user or by the initial bootstrapping process. This helps users review previous labeling decisions. Finally, at the end of each iteration, when the user is satisfied with her changes to the classification model, she can retrain the current classifier with the labeling controls (Figure 5.1 (i)). It includes buttons to label documents, retrain the classifier, and a bar that shows an estimation of the training progress. This estimation is based on the anticipated degree of change between the old and new model, a method from AL (see Chapter 2). It helps users decide how

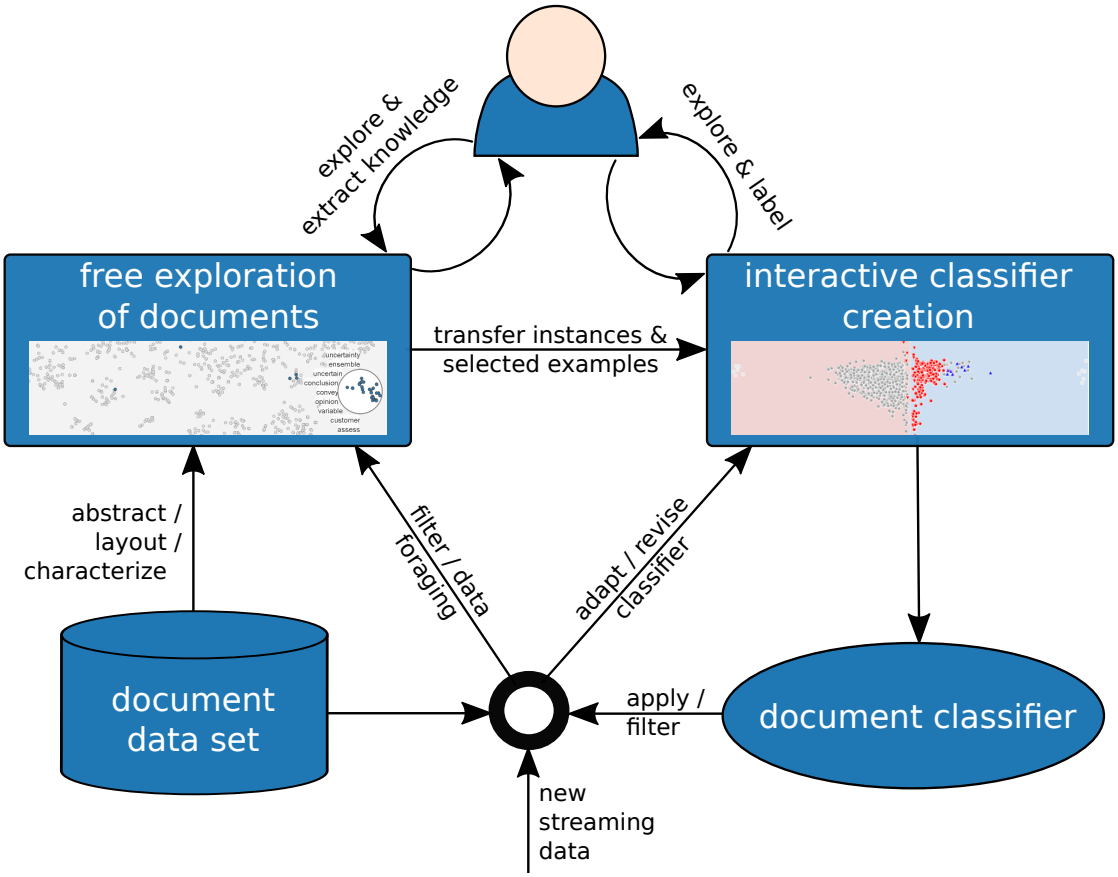


Figure 5.3 — The integration of our interactive classifier creation approach into an exploration environment for document data. The user bootstraps the process by selecting an initial set of relevant instances during exploration. During an interactive process, she then refines the classifier, and finally decides when to stop and store the classification model. The classifier can then be applied for data foraging [Pirolli and Card, 2005] on the original document data or on additional streaming data.

strongly a labeling action is going to change the following classification model in comparison to the current one.

After multiple iterations of this process, the user can finally decide when the classifier's performance is satisfactory. This decision is made based on a thorough exploration of the data set, which happens during the classifier creation process, and the exploration of the results and model characteristics of the current classification model. The final classifier can then be stored for further analyses, or applied to other, potentially dynamic data. The process of creating a classifier, including the

exploration of a data set with respect to a certain aspect, and the close exploration of the classification model, constitutes an exchange of information and knowledge between users and the machine. The user, on the one hand, decides about the aspect according to which the final classifier should classify documents. However, often, the aspect the user is interested in is initially a rather murky concept without definitive criteria. During classifier creation, the classification algorithm yields new cases of documents that it finds hard to classify, providing users with those border line cases that are particularly conducive to many forms of sensemaking (see Chapter 2). Also, reviewing and exploring confident classifications, and model properties provides users with new information about the murky classification criteria, and helps them to solidify their information need. The classification algorithm, on the other hand, gets direct feedback for its classification decisions. In particular, it is provided with definitive labels for particularly hard examples, allowing it to quickly learn new aspects of the classification problem, and to improve its performance at a fast pace.

The current stand-alone implementation of this approach has been created to provide a prove of concept, and to allow for an evaluation of the approach. However, in order to fully develop its analytic capabilities in conjunction with further exploratory methods, our approach is ultimately designed to be integrated into a larger text VA framework. Based on the previous exploration approach presented in Chapter 4, we have sketched such a framework, which is depicted in Figure 5.3. During exploration of the data set, the user gets interested in a particular topic, and decides to train a classifier for more in-depth exploration and analysis. She bootstraps the process by selecting a set of initial examples for classification. After the document classifier is successfully created, the user has significantly enlarged her knowledge about the topic. And the classifier has learned to distinguish between relevant and irrelevant documents with respect to this topic. The final classifier is then applicable on the original data set, for further in-depth exploration in a suitable VA environment or to new streaming data. The classifier is then an automatic means to perform efficient data foraging Pirolli and Card [2005]. If the user's trust in the classifier diminishes over time, e.g., due to a shift in the topics composition, the classifier can be adapted or revised any time based on a new or updated document data base.

5.4 Application to Scientific Literature

In this section, we describe a fictional training session that happens during the exploration of a scientific literature data set. The corpus contains all IEEE Vis(Week) full papers from 1999 to 2011, resulting in 1,329 single documents that contain 8,205,535 tokens and 56,622 types. To demonstrate the effectiveness of the

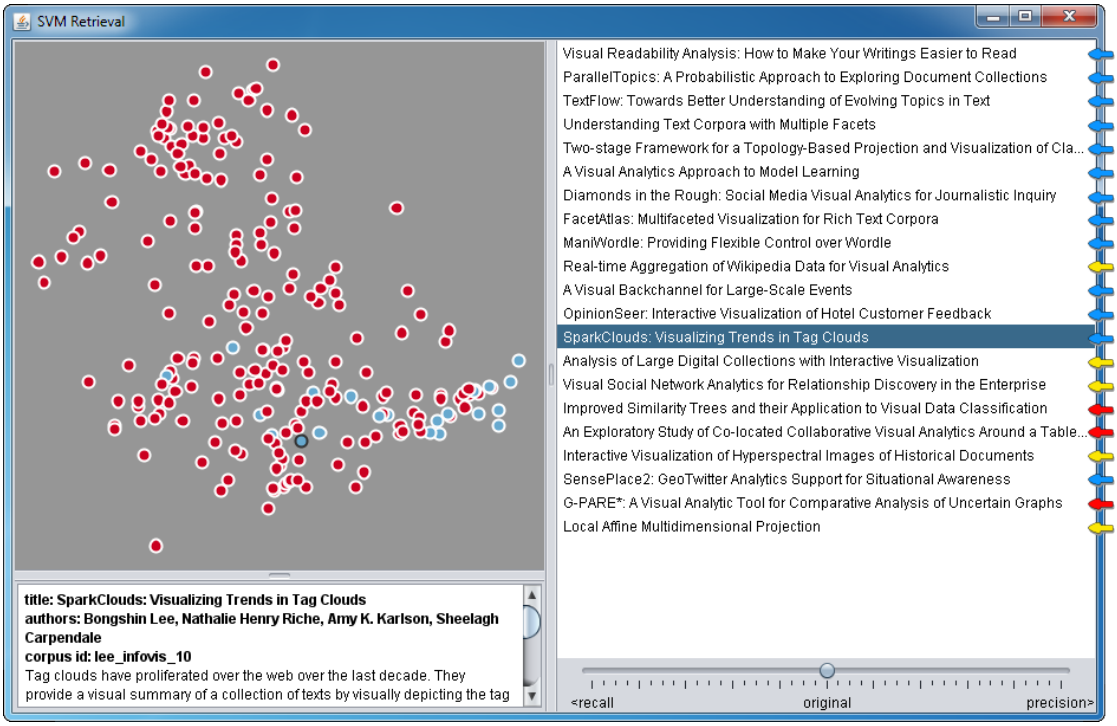


Figure 5.4 — Interactive exploration of the classification results. The small arrows at the right indicate the relevance of the retrieved documents according to the user's judgment. Blue arrows indicate relevant documents, yellow ones refer to border line documents that are partially relevant, and red arrows mark documents that are classification errors.

classifier, we split the set of publications, and use the ones from 1999 to 2009 (83% of the data set) for classifier creation. The remaining 17% of documents will be later used as a test set for the final classifier. To apply the created classifier to this held out data set, we have created a small application (see Figure 5.4) that shows the data set based on the LSP algorithm [Paulovich et al., 2008]. In the resulting spatialization, positively classified instances are colored blue, while negatively classified ones are colored red. The positive documents are also contained in a list next to the spatialization that is sorted according to classification confidence. A text view is included into the application that allows to view the content of selected documents. In addition, precision and recall can be adapted interactively, increasing or decreasing the number of positively classified documents by shifting the decision boundary of the classifier without changing its orientation in the dictionary space. In the following section, we report an example usage session of our approach based on a data set of scientific literature documents.

Figure 5.5 — The initial classifier configuration after bootstrapping, including the titles of the positive training documents in the white box (copied from the content view for this figure). All documents classified as positive are shown in the right, blue region. Negative documents are depicted in the left, red region.

5.4.1 Creating the Classifier

A person with an NLP background that has just started to delve into visualization realizes during exploration of the Vis(Week) data, that there is a host of text visualization publications. She tries to learn more, but finds that the publications about this topic are very diverse and not that easy to find. She thus decides to create a classifier to be able to explore this topic in more depth, and to automatically identify text visualization publications from future conferences and journals. After loading the data set into the system, the user bootstraps an initial classifier by using the very general query "NLP." This yields only a small number of six documents, resulting in the initial configuration depicted in Figure 5.5. It depicts all documents labeled positively by the bootstrapping procedure within the black rectangle. The documents' titles, as displayed in the content view, are shown in the white box. To keep the number of labeled positive and negative instances in balance, which is important for the SVM training algorithm, the same number of random negative examples is selected. They can be seen on the left side of Figure 5.5. As a first step, the user skims through the titles of the selected positive documents. For most of them, it becomes obvious that the respective papers are relevant to her. She reads over the abstracts of some documents to decide about their relevance. Doing this gives her additional insights into different aspects of the visual text analysis topic.

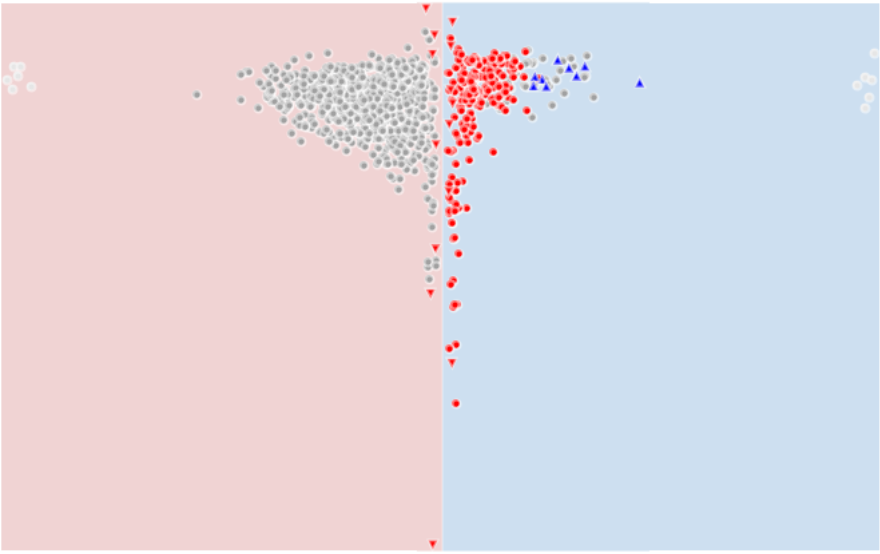


Figure 5.6 — Preview of the new classifier configuration at the end of the first training iteration. Documents colored red are the ones that are going to change sides (from blue to red) if a new classification model is trained.

She finally decides that all returned documents are relevant to her, and, after a glance at the negative examples, finds that they are correct, too. She then continues to explore unlabeled instances close to the decision boundary with the lens. This yields only negative examples, which she labels accordingly. Unfortunately, the ensuing imbalance of positive and negative examples in the training set produces a preview that classifies all unlabeled instances as negative. To counterbalance this situation, she starts exploring the far positive side of the space, close to the initial positive examples. Here, she quickly identifies a number of documents she decides to label as positive. The resulting preview is depicted in Figure 5.6. She is satisfied with this result and invokes a new training iteration.

After the training (see Figure 5.7), she immediately starts exploring the new model by inspecting the terms with high weights, as depicted in Figure 5.8. It contains many terms that were to be expected, such as *topics*, and *document*. One term that attracts her attention, and that she is unfamiliar with is *wordle*. Clicking on the terms selects all three documents in the data set that contain it. The user thus quickly identifies a publication that contains *wordle* in its title. From its abstract, she learns that *wordle* is a popular algorithm and internet platform to create visually appealing word clouds from texts. One of the three documents is still unlabeled. After a brief review, she labels it as a positive example. To further improve the classifier, the user again turns to labeling candidates in the area close

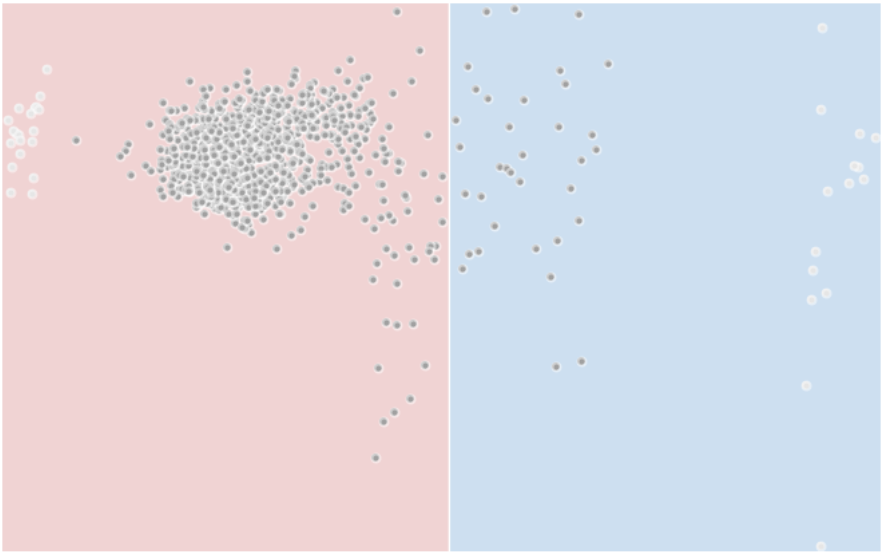


Figure 5.7 — Configuration of the classifier after the first iteration. Due to the labeling and subsequent training, a considerable number of documents that were classified as positive have now moved to the negative side.

to the decision boundary. She explores the documents in this area with the term lens and by hovering over single documents, and finds that she is satisfied with the current situation in the vicinity of the decision boundary. Then, in the cluster view, four tightly clustered documents attract her attention. The lens, as depicted in Figure 5.9, indicates that the focused papers are indeed all relevant. Looking at the respective titles and abstracts confirms her assumption, and she labels all four of them accordingly. The preview now shows that some negatively classified documents would turn out positive after a model update. She explores them briefly with the lens and by skimming through some titles and abstracts and decides that she is satisfied with the preview. After triggering the update, and some further exploration, she finds that she is generally happy with the model, but that it needs some fine tuning. To do this, she goes through three additional iterations, each time concentrating on the documents close to the decision boundary and labeling a number of additional ones. Finally, she is happy with the result and decides to stop the classifier building process.

5.4.2 Applying the Classifier

Now, after creating the classifier, the user wants to apply it to a newly published set of visualization literature. She wants to gauge the performance of the classifier on that set to see if she is satisfied with it, or if she has to revise it. She loads the

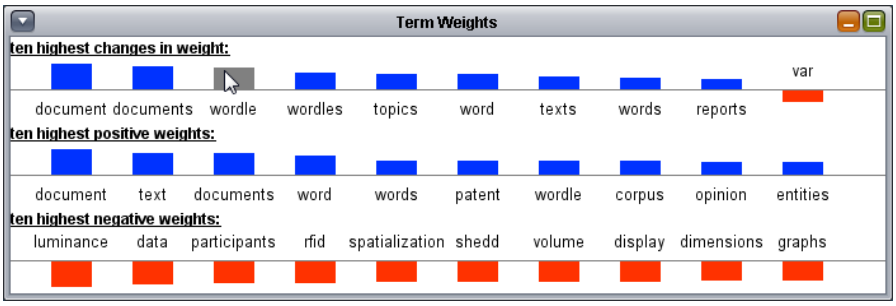


Figure 5.8 — Term weights during the second iteration. This round increased the positive weight of terms such as "document(s)", as can be seen in the "ten highest changes" chart. Below, the ten highest positive and ten highest negative terms of the current model are shown.

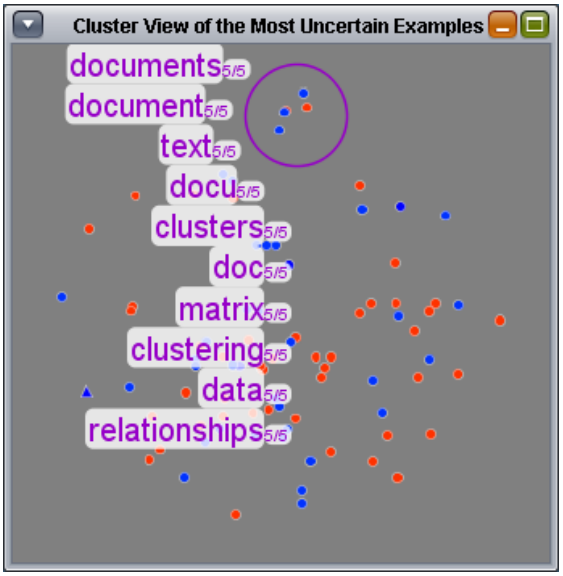


Figure 5.9 — Exploration of documents close to the hyperplane during the second iteration. By using the lens, the most frequent terms of the focused documents are shown. The lens helps to quickly explore the space and identify good labeling candidates.

classifier and the data set into the classification application, and starts exploring the results (see Figure 5.4). Inspecting the 2D map of the document set on the left side of the application, she sees that all positively classified documents, colored in blue, mostly occupy the lower left corner of the space. She briefly explores negatively classified documents particularly close to positively classified ones by going through their abstracts in the text view below the 2D map. As she cannot find any classification errors that way, she moves the recall/precision slider on the right lower side of the application window to the left. This causes more documents on the negative side of the decision boundary, but close to it, to be classified as positive. She skims these new texts that emerge in the list of positive results on the right upper side of the window. As she thus cannot find any missed texts either, she moves the slider back to the middle position and starts reviewing the positive results in the list. The colored arrows at the right side of Figure 5.4 indicate her judgment about class membership. Blue arrows mark relevant documents, yellow ones mark borderline documents that are partially relevant, and red arrows mark documents that are classification errors. The user thus identifies roughly two thirds of the documents as correct results (blue). From the remaining one third, three documents are clear errors (red). For the others (yellow), she discovers that most of them contain relevant information, although some of them cannot strictly be considered text visualization articles. Based on these results, the user decides that the performance of the classifier is sufficient, and that no adaption is needed.

5.5 Evaluation and Discussion

In this section, we report on an evaluation that we conducted to gain insights into whether training a classifier interactively produces satisfactory results with respect to classification performance. In the context of this thesis, we particularly focus on whether free exploration and choice of documents to label helps users create text classifiers. For this reason, we only discuss the relevant parts of the entire evaluation, and refer the reader to Heimerl et al. [2012b] for a more comprehensive evaluation of interactive visual classifier creation.

5.5.1 Data Sets and Procedure

With this evaluation, we aim at quantifying the users' success in creating classifiers based on the classification performance of the resulting model. For this, we need data sets with gold labels that allow to objectively determine the classifier's accuracy. We have thus decided to base the user evaluation sessions on two popular benchmark data sets from the NLP community. The first one is the 20 newsgroups collection (20ng), a set of 19,000 postings from twenty different

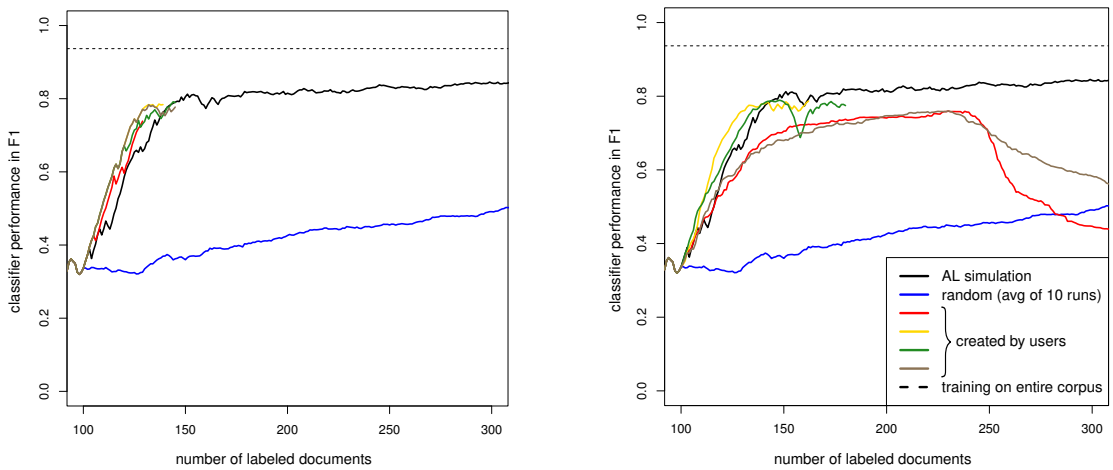
thematic groups from the usenet [Lang, 1995]. As the example search problem for this corpus, we decided to use all posting from groups starting with *comp.* as the positive examples (about 3,900), and all others as negative examples. This contains computer related posts including conversations about computer networks and various types of hardware. The second data set is a subset of 12,000 articles of the Reuters RCV1 collection [Rose et al., 2002] that contains newswire texts from 1996. All texts are labeled according to multiple categories, including their topic. We decided to use all sports news as the positive set, and the rest of the news as the negative one. The challenge with using those gold labels lies in the communication of these labeling tasks to users. To keep results of different users comparable, we have communicated the concept of the positive labels as clearly as possible. Nevertheless, according to [Saraiya et al., 2006], we have to expect negative effects from artificial tasks on data that the user is not tied to in any way. There is a third corpus that we use for an introduction to the systems before the user sessions. This is the Vis(Week) abstract data set, with around 1,200 abstracts from conference publications. As there are no gold labels for this corpus, the performance of the resulting classifiers cannot be measured afterwards. The example search task was the same as the one in the use case from Section 5.4, i.e., to find all visual text analysis publications in the corpus.

In the original evaluation [Heimerl et al., 2012b], we compared three different systems, but in the context of this thesis we will discuss the results for two of them. One is the *interactive system*, which is the one depicted in Figure 5.1. The second one is a restricted version that looks similar to the first one and allows free exploration of the data set. However, rather than allowing the user to decide which instances to label, one instance is selected during each iteration that the user has to label. The instance is selected according to uncertainty sampling, a technique from AL (see Chapter 2). This approach is called the *restricted system*. We chose these two, because, in the context of this thesis, we want to look at how exploration helps to determine the right documents to label as well as to steer classifier evolution. The goal of this user experiment is not to beat traditional AL with our interactive and explorative approach, but rather to learn whether labeling during exploration is effective for creating classifiers. For this, traditional AL that is typically used in very controlled environments does serve as a point of comparison to generate insights and spark discussions. In the original evaluation sessions, we had every participant use two out of the originally three methods for classifier creation. We permuted the data sets and systems to assure that each user gets two different systems and two different data sets to avoid learning effects. In addition, systems as well as data sets were presented in different orders to eliminate any influencing effects on each other. The sessions were conducted according to the think aloud

paradigm, asking participants to voice any of their thoughts, which we recorded on paper.

Twelve participants overall took part in the sessions. With our, originally, three systems and two data set, this means that we had four users for each system/data set combination. All of the participants were PhD students from our visualization department. They were between 35 and 55 years old. One of them was female, the rest was male. We had each of the participants complete two classifier creation sessions that together took about 60 to 90 minutes including introductions and discussions in the end. Overall, the procedure for each participant was as follows:

1. *introductory remarks*
   General instructions about the evaluation procedure.
2. *Ishihara color plate test*
   Test participants for any color vision deficiencies.
3. *introduction to the first system*
   Brief tutorial to the system. After that, participants could play with the first system and the Vis(Week) corpus until they wanted to switch to the actual data set.
4. *classifier creation session with first system & data set*
   Limited to 15 minutes, but participants were allows to stop any time before that.
5. *questionnaire about the first session*
   Contains questions about the first system and task.
6. *introduction to the second system*
   see 3)
7. *classifier creation session with second system & data set*
   see 4)
8. *questionnaire about the second session*
   see 5)
9. *general questionnaire*
   Collect personal information about the participants, such as age and their expertise in various field.
10. *open discussion*
    Discussion and feedback round with the participants.



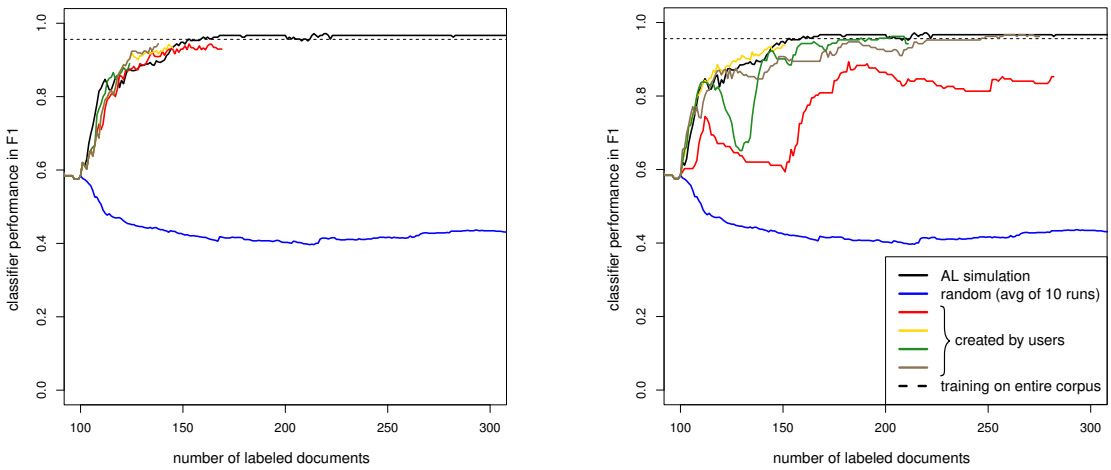
(a) Results for the restricted system (b) Results for the interactive system

Figure 5.10 — Classifier performance in F_1 (y-axis) over the number of labeled documents (x-axis) for the 20ng data set.

5.5.2 Classifier Performance

The bootstrapping strategy of the prototype is based on keyword queries, and includes a random element in that negative examples are randomly selected each time a query is executed. However, changing initial training sets would make the classifiers created by the users hard to compare, as each user started with a different initial configuration. For this reason, we decided to execute an initial queries once for each data set and store the results. This training data could then be loaded each at the beginning of each session. The queries we used for our three data set were "computers network motherboard graphics" for 20ng, "sports baseball basketball tennis game" for RCV1, and "text" for the Vis(Week) data. In addition, we used 20% held out data for each of the corpora to measure the performance of the classifier that was trained by exploring and labeling part of the remaining 80% of the data set.

Figure 5.10 depicts the resulting classifier learning curves that plot performance values measured in F_1 on the held out data against the number of documents labeled. Figure 5.11 shows the same plots for the RCV1 data. All diagrams contain a dashed lined, indicating the performance of a classifier trained on the entire 80% training portion with gold labels. The user generated performance lines are colored red, yellow, green, and brown. Each plot further contains a learning curve generated by randomly sampling documents from the training portion, and using their gold labels for training. This simulates a "dumb" labeler that has no particular strategy to select documents for labeling, and serves as the base line in



(a) Results for the restricted system (b) Results for the interactive system

Figure 5.11 — Classifier performance in F_1 (y-axis) over the number of labeled documents (x-axis) for the RCV1 data set.

this case. As the random sampling process can lead to very unstable curves, we have repeated the process 10 times, and depict the average of those runs in the plots. The black curves are simulated AL, i.e., AL with the gold labels and without any human intervention, which represents the "perfect" labeler in this scenario. For random sampling as well as simulated AL, we used to same initial training sets as for the user sessions, to keep the results comparable. As this set contains 100 instances, all the plots start with 100 on their horizontal axis. Figure 5.11 shows one phenomenon that deserves mention. Some of the learning curves pass the dashed line, which means that they achieve higher performance with fewer labeled documents. The effect is well known and has been described by Schohn and Cohn [2000].

5.5.3 Questionnaires

The questionnaires that we asked participants to fill out after the individual sessions contained the following questions, most of which were to be answered on a seven point Lickert scale:

- All six questions from NASA-TLX [Hart and Stavenland, 1988] to determine the user's task load.
- How would they rate their trust in the classifier.

- Why did they stop building the classifier.
  (With several options including free text instead of a Lickert scale.)
- How they rate the usefulness of each view of Figure 5.1.
  (With one Lickert scale per view.)

The general questionnaire that participants were asked to fill out after both sessions were finished contained the following questions of which, again, most were to be answered on seven points Lickert scales:

- What their age was.
  (With no Lickert scale.)
- What their gender was.
  (With no Lickert scale.)
- Their expertise in web search engines.
- Their expertise in machine learning / classification.
- Their expertise in using interactive visualization.
- Their expertise in general data / information finding tasks.
- Which of the two methods they preferred.

5.6 Discussion of Results

In the general questionnaire, the users rated themselves on average as follows (a higher value means more experience):

- using web search engine: 6.15
- expertise in machine learning: 3.31
- experience with interactive visualization: 5.92
- experience with general search tasks: 5.46

The users' experience with interactive visualization was rather high, as was to be expected. This had the advantage that they were able to learn rather quickly how to use the systems. Some of them also had machine learning experience, some did not, which led to a lower average number.

Four users for each system / data set combination did not yield statistically significant numbers. We therefore base the discussion on the individual results achieved in the sessions, as depicted in Figures 5.10 and 5.11. The plots indicate that it is possible to create high performance text classifiers interactively. In fact, all of the methods, including the simulated AL curve, have approximately identical peak performances. However, the plots show some phenomena that deserve deeper discussion. One conspicuous fact when comparing the plots for the interactive method to that of the restricted methods is that users were able to label much more documents when working with the interactive method. The reason for this is twofold. Firstly, the restricted method constrains the number of instances labeled to just one which is determined by the system based on AL criteria. This means that users, unlike in the interactive sessions, cannot label larger numbers of documents during exploration. The incentive for some users to actually start exploring the document set in more depth, rather than just go with the one document that has been suggested by the system was also reduced by this. Secondly, retraining the classifier and re-calculating the visualization for it takes a couple of seconds. This introduced some amount of waiting time into the sessions with the restricted system. In addition, most of the participants did not stop prematurely, and continued labeling and exploring until the 15 minutes were used up. The labeling rates for the interactive sessions thus turned out to be significantly higher on average due to all of these reasons.

Despite the positive results, however, we can also see that users of the interactive method produced some suboptimal learning curves. This was naturally not the case for the restricted method, as the AL procedure chose documents that provided high potential for the development of the classifier during each iteration. The interactive method, however, introduced more playfulness into the classifier creation process. This could be observed during the sessions, and the participants mostly rated the interactive system as the one that did not get them bored as easily as the restricted one. As playfulness is an aspect that is conducive to the sensemaking processes according to Klein et al. [2007], this a positive outcome. Despite improving knowledge generation from the data, the desire of participants to try out certain actions sometimes lead to training documents that hurt classifier performance, as can be observed in the plots. However, the participants were also able to quickly assess disadvantageous labeling actions as such, and improve the situation again during subsequent iterations. This shows that despite the higher chance of things going wrong that comes with increased user control over the process, the knowledge gained during exploration helps to detect and solve such problems.

Another effect discernible from the plots is that training on the RCV1 data seems to be generally more successful compared to the 20ng corpus. Better performance results are achieved for most of the curves, for the simulated ones as well as for

the user generated ones. The reason for this seems to be, that the RCV1 data is curated and contains only documents with meaningful contents, while 20ng has been collected from internet newsgroups. Next to longer posts that contain text, it contains several very short documents with little discernible content, e.g., consisting only of smilies. This sometimes makes it very hard, or even impossible to discern the class of a document based on its content. The plots show that this seems to be hard for the classifier, as well as for human users. In particular, this is the reason why some of the curves pass the performance of the classifier trained on the entire data set. Carefully selecting instances seems to reduce noise in the training data, and improve overall performance. This can be achieved by human users, but also by the AL procedure used for simulating the AL run. The additional strain that this puts on users can also be seen in the results of the NASA TLX questions from the questionnaire. Here, the session involving the 20ng corpus are rated as causing much more mental stress compared to the RCV1 corpus. This hinders efficient exploration, as spurious data may lead to rejection or questioning of the current frame [Klein et al., 2007], triggering sensemaking processes that deal with how to handle this data. Thus, processing this data can take up a significant amount of time. For the RCV1 corpus, in contrast, exploration of relevant and non-relevant articles immediately provides users with useful insights into the various types of sports covered within the data set, and how the articles about each of them look like. Learning about these things help users immensely with deciding which documents to label.

Another effect that we saw during the evaluation sessions suggests that the interactive system needs sufficient user training before it can be used effectively. Even for visualization experts it took some time until they were proficient enough with the system to start building a classifier. We found that the training session before the real classifier building task was often not enough to convey sufficient knowledge about how to use the system. In conjunction with ambiguous data instances, this led to disadvantageous labeling actions that the users were not able to recover from. The red and brown curve in Figure 5.10 (b) are examples for this. While the visualization of the classifier state and the content view (Figure 5.1 (b) and (e)) were rated helpful by all of the participants in the questionnaires, the term lens and the term weights view (Figure 5.1 (c) and (f)) were not straightforward to use for some of them. We could also observe this during the sessions, e.g., when the term weights view led to large labeling actions based on the occurrence of one single term. Interestingly, participants that indicated that they have higher machine learning experience in the questionnaires did not produce better performing classifiers than other participants. This could indicate that the presented approach can be effectively used independently of the user's experience with classification, although the

number of participants in this experiment is too small to yield conclusive results about this.

To summarize, we can say that the results indicate that interactive classifier creation is suitable for users who are willing to invest some time into learning to use the system, and for whose workflows search and foraging [Pirolli and Card, 2005] tasks play an important role. However, due to the limited number of participants, and the complexity of the tasks of exploring a data set and labeling suitable documents, many of the results are still inconclusive. This calls for a more extensive study in the future that does not only focus on classifier performance, but looks at the entire process including sensemaking and exploration during the sessions in more detail.

5.7 Domain Extension

After these encouraging results, we decided to extend the approach for interactive text classifier creation to a different domain, to demonstrate its versatility. Together with fellow researchers who were working on the analysis of social media for situational awareness, we adapted the approach and applied it to microblog messages. They were especially interested in being able to accurately filter messages that describe certain events from microblog services. The message streams on microblog platforms, such as Twitter, has reached extremely high volumes. Furthermore, their dynamic, and up-to-date nature make it a great information source for situational awareness and decision making, e.g., during large events, or even natural disasters. For this, microblog analysis systems need methods to quickly and effectively process large amounts of very short, but highly noisy texts and gauge their relevance to a specific topic. In this context, simple keyword queries to filter messages are neither precise enough, due to the high noise level within these very short texts, nor scalable enough to be easily adapted to the dynamics of these information channels. Also, such application areas fulfill the criteria for potentially successful interactive classifier training that we discussed at the end of the previous section. As part of the ScatterBlogs2 system [Bosch et al., 2013] for microblog filter orchestration, we have developed an adapted version of the classifier creation approach. It is specifically designed for the specific requirements of such texts, and helps users create microblog filters based on SVMs for text classification.

Microblog messages are very short texts with up to 140 characters. To represent them adequately for classification, we have changed the underlying representation of documents from tf-idf vectors to Boolean vectors. This saves processing power and memory and renders the approach more scalable to the high level of noise, and huge messages sets that may even contain texts in many different languages. In

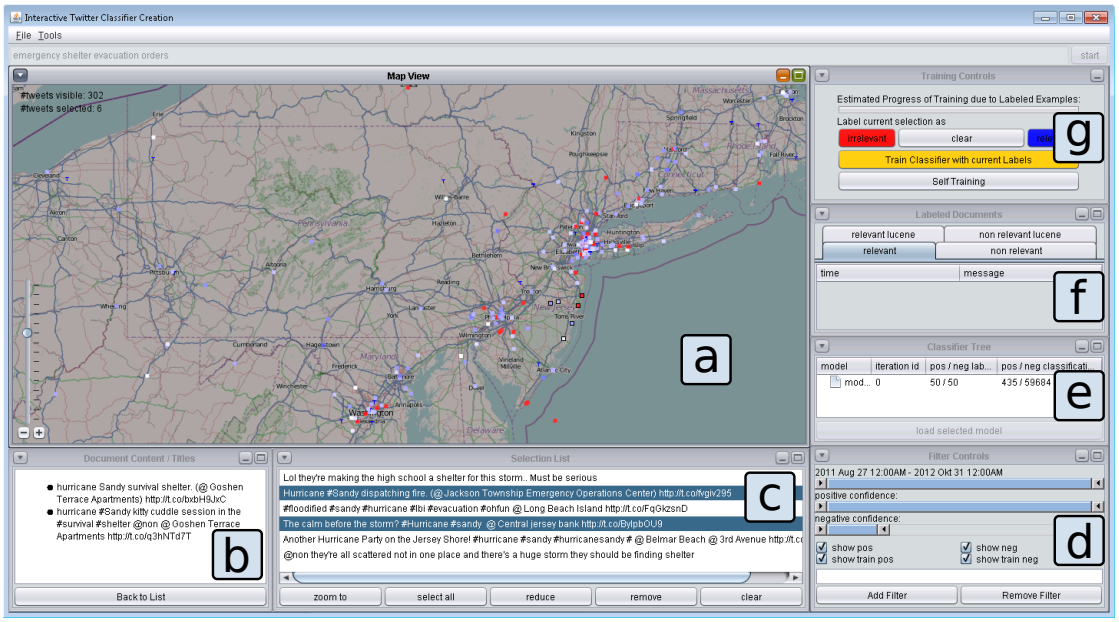


Figure 5.12 — The classifier creation desktop for microblog messages with the map view (a), the content view (b), the selection view (c), the filter controls (d), the classifier history (e), the labeled documents view (f), and the controls for the labeling functions (g).

addition, the highly dynamic nature and location dependence of term use [Thom et al., 2012b,a] suggests that idf is not an adequate measure for term popularity for this type of documents. We have further adapted the SVM framework that the approach is based on. It now not only supports linear kernels, which we use for other documents due to its fast training and classification times, and its high accuracy for classifying longer documents. For the microblog messages, we offer a second kernel function that users can choose, namely the string kernel [Lodhi et al., 2002]. It operates directly on the strings to compare two texts, and is able to deal with typos, slang, and other types of noise that microblog texts often contain. As it has the drawback of added computational complexity, and thus need more time for training and classification than linear kernels, we let the users decide what they think is adequate for their particular application. Figure 5.12 shows the adapted version of the system. It shares many properties with the previous system in Figure 5.1, but has received multiple changes that we discuss hereafter. The classifier is to be trained on a set of previously recorded messages that contain data from events that the user wishes to detect. This could, for example, be messages about roadblocks that people send via microblog, e.g., due to collapsed trees or electric poles during a heavy storm. Users record such a data set, and can then

load it into the system, explore the messages, and iteratively create a classifier that is able to detect them. The only condition our approach has for the corpus, is that all messages must contain a geo-location. However, this is typically the case for messages that are used for situational awareness and many areas of decision making in which geographic information plays a role.

Bootstrapping, here, is also done with a keyword query whose result can be explored by users and corrected if necessary. For this, the map view (Figure 5.12 (a)) that depicts each of the message at its assigned geo-location as rectangles can be used. The colors of the rectangles convey classification results and uncertainty. Messages right at the decision boundary of the SVM are colored white, those positively classified are colored blue, and those negatively classified are colored red. The intensity of the color is determined by the distance to the decision boundary of the SVM. In addition, messages labeled during the current training iteration change their shape to triangles and their color to the assigned label, while messages that are part of the training set are shaped like a capital T. User can explore this space by hovering single messages and by using the lens. The lens lists the most popular hash tags in the messages under it, thus using the categorizations that users give their messages for aggregation. Highlighted messages are listed in the content view (Figure 5.12 (b)) with their full contents. In addition, users can select multiple messages to add them to the selection view (Figure 5.12 (c)), which helps users to review their selection, and add or remove single messages if they wish to, e.g., before labeling the entire selection.

The filter controls (Fig. 5.12 (d)) offer further assistance in finding relevant messages. They include a range slider for constraining the publishing date and time of the messages displayed, e.g., to show only messages published during the known time frame of an event. A replay of the arriving messages can be achieved by constraining the creation time of displayed massages to a small timespan and then moving the slider. Thus messages pop up in the order they were published to trace, for example, the route of a storm. The map view in combination with the temporal filter enables users to find past events by their spatio-temporal extents. Two additional range sliders let users restrict the confidence range of positively and negatively classified messages on the map view. This makes it possible, e.g., to view only the messages classified with very low confidence by the current classifier for further inspection. The filter controls further allow to turn the display of positive, negative, and training examples on and off separately and to filter messages by keywords. This is useful, e.g., to find messages containing a certain hash tag or place name.

The modified approach has been implemented and used within a larger framework for microblog message filtering and filter orchestration. Bosch et al. [2013] describe this framework and provide multiple use cases that include the use of text classifiers created with the described approach. They successfully apply it to filter social

media data for situational awareness during events such as natural disasters and disease outbreaks.

5.8 Future Directions

In this chapter, we have presented a novel approach for the creation of text document classifiers that can be used during text exploration. It is designed for initially murky information needs that solidify during the process. This is achieved by facilitating free exploration of the text data set and the classifier state and performance to derive insights and knowledge about which instances to label. Thus users and the automatic classification algorithm learn from each other during an AL-based classifier creation loop. Especially the linguistic knowledge and sensemaking capabilities of humans help drive this process. We have presented a prototypical implementation and sketched its integration into a larger framework for text exploration. In addition, we have reported on an evaluation of the approach and its implementation in a user study that suggests its effectiveness under certain conditions. Finally, we have presented the adaption of the approach to the domain of microblog messages, with its special requirements. This shows the versatility of the interactive classifier creation method.

One area for future research endeavors into visual interactive classification methods is the design of provenance visualizations. It would be great if users that applied the classifier on a document set could view visualizations that convey the reason the classifier has made a certain decision. This would help to decide whether to trust a classifier, or quickly identify faulty classifications. Users could then interactively modify the basis for the decision, incorporating only sources that they trust or deem adequate. Creating such methods would require research into visualization schemes that accurately convey this information, and suitable interaction techniques to allow users to interact with these visualizations. In addition, algorithms to create and efficiently store the respective information from the training sets would also have to be devised. Another future area of research is the extension of the approach to methods that help with other tasks than search and foraging. This could involve more complex classification tasks, such as multi-class problems, but also other tasks, especially in the domain of scientific literature, such as community detection, success prediction, or review assignment.

CHAPTER 6

Interactive Visual Feature Exploration

This chapter contains previously published material from the following publications:

F. Heimerl, C. Jochim, S. Koch, and T. Ertl. Featureforge: A novel tool for visually supported feature engineering and corpus revision. In *International Conference on Computational Linguistics (COLING) Posters*, pages 461–470, 2012a

In the previous chapter, we have looked at an approach designed to let users explore the state of a classifier with respect to a specific data set during exploration. It is designed to help users create and improve classifiers while they are learning about the data set and the specific classification problem. The classifier can later be applied for search and foraging tasks. In this chapter, we aim to support classifier creation with a different goal, but through much of the same means, i.e., exploring its state and the underlying data set. However, in contrast to the previous approach, the one presented in this chapter is designed for NLP specialists rather than domain experts or users interested in a particular data set. FeatureForge has been developed as a collaborative project with NLP researchers, and is designed

to help them with making sense and exploring their linguistic data. The focus of the described method lies particularly on the labels of the underlying data, and the feature representations of the linguistic entities to classify. Thus, the approach described in this chapter is an example of approaches designed according to the specifications of the lower part of the two level model for language data analysis (see Chapter 2). The resulting classifiers and language processing methods can then be used for text mining, or, more importantly for us as part of a VA approach to help domain experts explore and analyze their data.

6.1 Motivation

Supervised machine learning methods provide state-of-the-art performance in many areas of NLP. Researchers who want to advance the state-of-the-art either deal with the development of new classification methods that better model the linguistic data, or improve the training data itself by developing clean and accurately labeled corpora and feature definitions that allow for an effective vector representation of linguistic entities and provide as much discriminatory information as possible for accurate classification. The method presented here aims to support researchers in NLP with the latter task. In the following, we discuss the prototype of an interactive software system that helps researchers with spotting problems in their vector representations, as well as with the gold labels of their corpora. It is based on interactive visualizations of the data, the classification model, and the feature representations to support the sensemaking processes of NLP experts. Despite many powerful methods that have been devised to automatically select useful and discriminatory features (see Somol et al. [2010]) from a large set of possible representations, the linguistic skills of experts are still of utmost importance in the process, as human knowledge and language skills are needed to create useful features in the first place. For this reason, Guyon and Elisseeff [2003] identify the feature definition step as the one that should be tackled first when trying to improve classification performance. In order to support free exploration of all relevant information and help NLP experts gain knowledge about their data that helps them to develop new features, we combine various visual interactive data aggregation and filtering methods. These can be used for exploration, to highlight relevant aspects of the data, and to gain insights into the classification models. In addition, visualizations can be steered and controlled through user interaction. The presented approach combines interactive visualization techniques with unsupervised clustering methods to effectively support researchers with the task of refining vector representations by creating new features, as well as finding systematic annotation errors in the corpus. It provides information about the effectiveness and discriminatory power [Somol et al., 2010] of feature representations.

This is achieved by displaying the current state of the classifier based on these features, in combination with a hierarchical clustering of the instance space to help users identify problematic instances. NLP experts and feature engineers can examine such instances, derive new features, implement and test them. When adding a new feature, the system updates all views instantaneously and thus gives much richer feedback about their impact than just the bare performance numbers of the resulting classifier. Apart from deriving new features, mislabeled instances can help update annotation guidelines by refining or adding annotation rules. Some of the mislabeled instances can additionally be used as examples for annotation guidelines in order to convey labeling rules more effectively. This closes the loop between corpora that are developed and labeled by linguistic experts, and the statistical machine learning algorithms to develop automatic language processing methods.

6.2 Related Work

The contributions to the ACL Feature Engineering Symposium [Ringger, 2005] show that feature engineering plays an important role for many NLP tasks. Carefully and resourcefully engineered features are able to significantly improve classification performance. NLP problems tackled by the symposium contributions include, but are not limited to, text segmentation [Kauchak and Chen, 2005], shallow semantic parsing [Moschitti et al., 2005] and recognition of temporal expressions [Adafre and de Rijke, 2005]. There is evidence [Scott and Matwin, 1999] that domain-specific feature engineering is also beneficial for a task such as text classification, where the bag-of-words model is established and shows good performance in many situations. Berend and Farkas [2010] present a set of new and effective features for automatic key phrase extraction from scientific papers. Those examples all show the importance of linguistic feature creation for statistical NLP applications. Kobdani et al. [2010] identify the creation of feature representations for linguistic instances as the most crucial and costly step in the process of building an NLP application. They offer support for this step by proposing a feature definition system based on a data base model that offers a lot of flexibility by exploiting the capabilities of the SQL language.

There are approaches to feature space examination that build on visualization to give users insight into the usefulness and discriminatory power of feature representations. Kienreich and Seifert [2012] apply matrix reordering algorithms on document-term matrices and classify and discuss typical emerging patterns. With the help of those patterns, users are able to judge the usefulness of certain terms for the discrimination of classes. Schreck and Keim [2006] create a 2D map of an instance space produced by a feature extractor for multimedia data sets. Based on their

approach, users can evaluate the effectiveness of automatic feature extractors for multimedia data sets. They hypothesize that the uniformity of distances between instances in different clusters and low variance between instances in one cluster correlates with the discriminatory power of a feature representation and provide visualization techniques that help users assess both characteristics. Dolfing [2007] developed a VA approach for feature engineering based on an interactive clustering algorithm that, unlike our approach, directly incorporates user judgment in its clustering decisions. The presented system is tailored to enable feature space analysis for optical character recognition. Partially similar to our approach, it includes a component that displays a classifier's state by plotting the distribution of data around its decision boundary. The popular machine learning system WEKA [Hall et al., 2009] offers user interfaces for clustering and classification including visualization to support the user with understanding the data better. WEKA also includes methods for feature selection. Contrary to FeatureForge, however, it does not offer any linking between clustering and classification to support the feature engineering process and offers no solution for corpus improvement.

Feature engineering is related to feature selection, which comprises a number of techniques to automatically reduce the number of dimensions of instance vectors in order to either produce more compact representations, or optimize the representations for a certain classification task. Liu and Yu [2005] provide an overview of feature selection methods and classify them into different categories. Somol et al. [2010] developed the Feature Selection Toolbox, a library for feature selection. Recently, Brooks et al. [2015] have created another VA approach for feature development. In their system, features are based on the combination of terms from texts. They support feature engineers with their task by providing systematic overviews of errors by grouping similar misclassified instances. Stoffel et al. [2015] present a visual feature exploration approach that provides overview visualization of text features. It helps users judge the importance of single features with respect to the classification task, and the frequency with which features were involved in misclassifications.

6.3 Interactive Feature Exploration

In this section we present FeatureForge, our approach for classifier exploration that is designed to help scrutinize features and gold labels of language data corpora. The first part of this section presents the FeatureForge desktop, depicted in Figure 6.1 by going through the views and explaining their purpose one by one. It describes what information is displayed, how it is presented to the users, and how it helps to explore the data set and features. All views are connected through brushing and linking, and thus selections and highlights are propagated from one view to

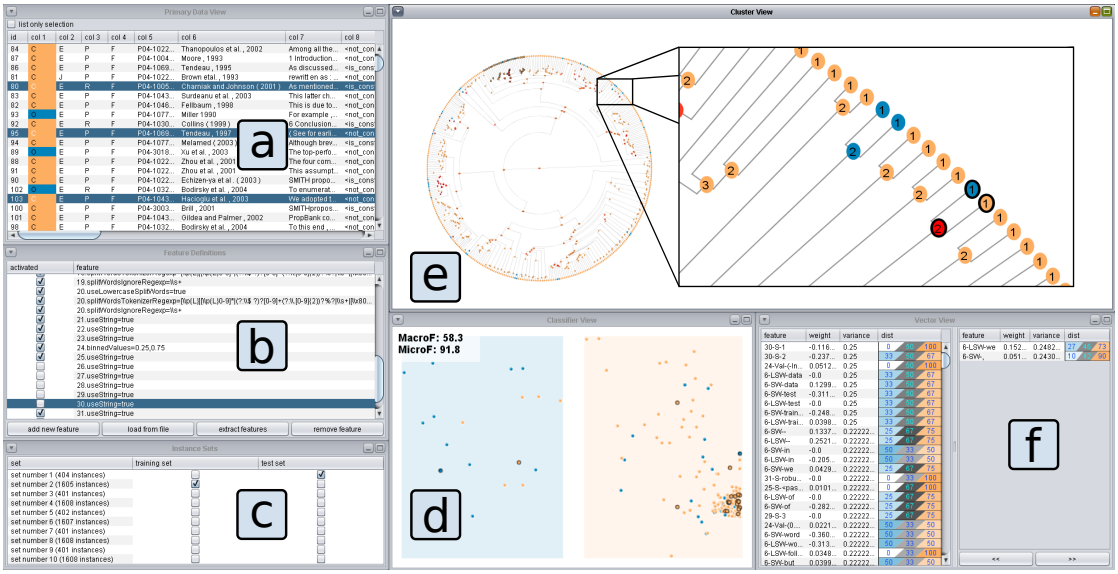


Figure 6.1 — The FeatureForge desktop with (a) the primary data view, (b) the feature definition view, (c) the instance set view, (d) the classifier view, (e) the cluster view, and (f) the vector set view. The screen shot shows the cluster tree zoomed out and part of it enlarged for illustration.

all others. FeatureForge currently integrates the feature definition language of the *ColumnDataClassifier* which is part of the Stanford classifier suite [Manning and Klein, 2003] and makes them available to users. This particular language is restricted to Boolean feature definitions, resulting in Boolean instance vectors. The prototype supports the column-based file format of the *ColumnDataClassifier* for the primary data of instances. Primary data is the data on which feature definitions are based, and which has been extracted from the linguistic entities through various preprocessing steps. They can contain, e.g., parses of the sentences of a text, named entities recognized, or resolved co-references. Users are able to load an arbitrary number of files containing instance sets into the instance sets view (Figure 6.1 (c)).

The primary data view (Figure 6.1 (a)) displays a table containing the instances from the instance sets loaded by the user. Alternatively, the view can be configured to show only the currently selected instances by checking the check box on top. Additionally, instances can be pinned to the top of this view to track them over multiple iterations. This is helpful, because after each iteration, the layout of the views changes, which can make it hard to spot critical instances from previous iterations. The first column contains a unique id for each item, the second one contains the gold label which is also indicated by color. Users can freely select a

color for each class in the data set. The current prototype supports only data with binary labels (we chose the colors blue and ocher for them in Figure 6.1). The primary data view contains all information available for each data item. Depending on the data, this can provide additional information, e.g., about the textual context of the instances.

The feature definition view (Figure 6.1 (b)) shows a list of the currently defined features and allows the user to add, remove, activate and deactivate features at any time. If a feature is removed or deactivated, the respective attribute is eliminated from the vector representation of each instance for classification. Feature definitions can also be loaded from file. After users have modified the list of feature definitions in this view, the feature representations for the currently loaded instance sets can be updated by clicking the *extract features* button.

The instance set view (Figure 6.1 (b)) holds a list of all the loaded instance sets. At least one training and one test set have to be selected as development set for feature exploration. The training set is used to train a classification model with one of the learning algorithms integrated into FeatureForge, while the test set is used for classifier evaluation and as a basis for the definition of new features.

The classifier view (Figure 6.1 (d)) shows the classification model with the test set to give users an impression of the classifier's state and performance. It is based on the 2D visualization from the previous Chapter 5, but it has been modified in some aspects that we will describe here. Currently, the classification methods integrated in the FeatureForge prototype are the *linear* and the *logistic* classifiers from the Stanford suite, and linear support vector machines based on the LibLinear library [Fan et al., 2008]. The decision boundary of a model is depicted as a white line separating the two instance half spaces. All instances of the test set are mapped on the 2D plane of the classifier view. For the arrangement along the horizontal axis the classifier confidence is used. Depending on the machine learning method, the confidence values are either the distance from the decision boundary for SVMs, or the probability estimate for the other two. The position along the vertical axis is determined by the first principal component of the 50 most uncertain instances on each side of the decision boundary, similar to Chapter 5. This keeps the fidelity of the 2D representations for the most uncertain documents highest. The vertical position of all other instances is determined based on the positions of their nearest neighbors among the uncertain instances, the same way as it is in Chapter 5. All instances from the test set in the classifier view are colored according to their gold label. This makes misclassification easy to spot.

Performance values of the depicted classifier are also included in the performance panel in the upper left corner of the classifier view. It shows the micro- and macro-F scores (see Chapter 2) of all depicted test set instances. The classifier view provides

insights into how the classification model treats the similar yet heterogeneously labeled instances. Such instances may reveal problems with features or label quality, as described later. If, e.g., the instances are scattered near the decision boundary, classification confidence for them is low which could indicate problems of data sparseness or missing discriminatory information. If they are classified with high confidence, this could be indicative of systematic labeling errors. In case they get assigned correct classes by the classifier, no intervention is necessary to improve performance. However, additional features can still help to increase classification confidence.

The cluster view (Figure 6.1 (e)) displays a hierarchical cluster tree of the test set computed by an agglomorative clustering algorithm based on Euclidean distance between the instances. We use Ward's linkage [Ward, 1963] as the cluster similarity measure. During each iteration, it joins the two clusters that result in the minimal increase of the residual sum of squares with respect to the cluster centroids. Once training and test set have been specified, and an initial feature set is available, the cluster view displays a hierarchical clustering of the test set. A popular visualization method for hierarchical clusters are dendrograms [Manning et al., 2008; Seo and Shneiderman, 2002].

We have decided to use a radial dendrogram to visualize the clustering of the test instances. The advantage of the radial layout is that it provides more space for the growing number of nodes towards the perimeter of the circle thus providing a better overview of the cluster tree. Instances with identical feature representations are joined before the clustering is started and form one common leaf node in the tree.

Each node is labeled with the number of instances that the corresponding cluster contains. The color of the nodes is chosen according to the signed homogeneity value, which we define as $(\frac{2 \cdot |c_1|}{|c_1|+|c_2|} - 1)$, where c_1 and c_2 are the sets of instances of the two classes in the cluster. The color of the node is the color of c_1 if signed homogeneity is $+1$, and the color of c_2 if its value is -1. If the homogeneity value reaches 0, the node color is red. For values in between ± 1 and 0, the color is interpolated between red and the color of c_1 or the color of c_2, respectively. This allows users to quickly identify heterogeneously labeled instances within one cluster.

The purpose of the cluster view is to guide users' attention to nodes that are good candidates for closer inspection. If a node in the cluster view is selected or hovered, all child nodes are highlighted and a selection event is triggered for all instances contained in the respective cluster. We hypothesize that instances very similar with respect to their vector representations having heterogeneous gold labels are a symptom of missing discriminatory information or a result of mislabelings. Such

nodes have higher heterogeneity and lie close to the perimeter of the radial tree. This means that the better the discriminatory power of the vector representation, the more red nodes, which indicate high heterogeneity, concentrate near the center of the radial tree. With respect to visual scalability, the cluster view has some limitations with its current design. It can accommodate up to 1,000 instances currently, which, given the fact that it only contains the test set of the data should be sufficient for many data sets. However, merging uninteresting subtrees into single nodes that can be explored on demand is a straightforward way to improve scalability and accommodate even very large test data sets. Uninteresting clusters are mostly those that are homogeneous with respect to their labels and feature representations.

The vector set view (Figure 6.1 (f)) provides information about the attributes of the selected instances. Selecting sets of instances can be done by either selecting a node in the cluster view or multiple instances in the classifier view. The vector view is subsequently filled with a list of the currently extracted dimensions[1] based on the active feature definitions. Each dimension occupies a row of the table that contains the following information, in order of the columns: textual description, weights of the classification model, variance of values, occurrences of attributes. The last column is partitioned into three blocks. At the center, the fraction of instances that have the respective attribute is displayed as a numerical value and with color being interpolated between white (0%) and black (100%). The left block displays the fraction of them that belong to c_1, and the right block those that belong to c_2. Both blocks are interpolated between white and the color of the respective class.

By using the buttons on the bottom right, selected dimensions can be moved to the table on the right. If it contains dimensions, the left table is restricted to the set of instances that have the respective attributes, i.e., it shows the conditional distribution of attributes. The vector set view characterizes selected instances in terms of the defined features by showing in what dimensions they differ and how the attributes are distributed over the instances. This provides hints about how well certain attributes are an indicator for a class in the selected set. By using the second table, interdependencies of the attributes can be explored, thus offering more fine-grained information about the distribution of attributes.

The FeatureForge approach provides multiple linked views for effective data exploration. Through its different views, it links clustering and classification on a visual level and through interaction, but not algorithmically. While this type of linkage

[1] We use the term *feature* for the user defined units to describe the data, e.g., a bag-of-words representation for a sentence. *Dimension* describes a position of the resulting vector and *attribute* denotes the respective instance property. Examples for dimensions / attributes are, e.g., single terms for a bag-of-words feature.

is designed to help users explore the properties of a feature set and learn about its problems, we cannot guarantee that all feature or labeling problems can be identified with this approach. However, comprehensively identifying all problems is not possible automatically, and can only be based on human knowledge and experience. For this reason, we have designed FeatureForge, to support experts with the art of feature engineering.

6.4 Case Study

For our case study we concentrate on citation classification [Teufel et al., 2006]. Each citation, extracted from scientific texts, constitutes an instance for classification. We report problems with the data and its features we discovered during an example session and sketch solutions for those findings. The data set comes from the ACL Anthology Reference Corpus [Bird et al., 2008] and has been annotated by NLP students. We use this corpus because it has been developed by our NLP partners, including the annotation guidelines, and we are familiar with it. Meanwhile, the techniques and the data set from this use case have been published in a paper on citation classification by our project partners [Jochim and Schütze, 2012]. The citations are labeled according to the scheme of Moravcsik and Murugesan [1975]. It comprises four facets with two possible labels for each, resulting in four binary classification problems. One strength of this scheme lies in the possibility to combine facets for finer grained labels. For the purpose of this illustration we focus on the facet *conceptual vs. operational*. *Conceptual* citations contain an idea or concept relevant to the citing paper, while *operational* ones contain a tool or programming library that the citing paper uses. The complete data set contains 2009 citations, which we split into a training and a test set (80%/20%) on the granularity level of the documents containing the citations, resulting in 1605 training and 404 test instances. For this use case, the Stanford logistic classifier is used for classification. We load an initial feature set containing 32 features that were previously created for the classification problem. Examples of features are a bag-of-words representation for the sentence containing the citation, the position of the citation in that sentence, whether the citation is a constituent of this sentence, and whether the sentence contains any named entity that denotes an NLP tool.

We activate all initial features and start exploring the cluster view. Here, we see a rather high number of citations in heterogeneous nodes on low variance levels. Browsing and comparing these instances in the classifier view, we realize that they are responsible for a large number of misclassifications, and that the classifier is not able to keep these instances apart due to their nearly identical feature vectors. Checking the primary data view, we realize that those highly similar citations occur within the same sentence in their source documents. The fact that one of

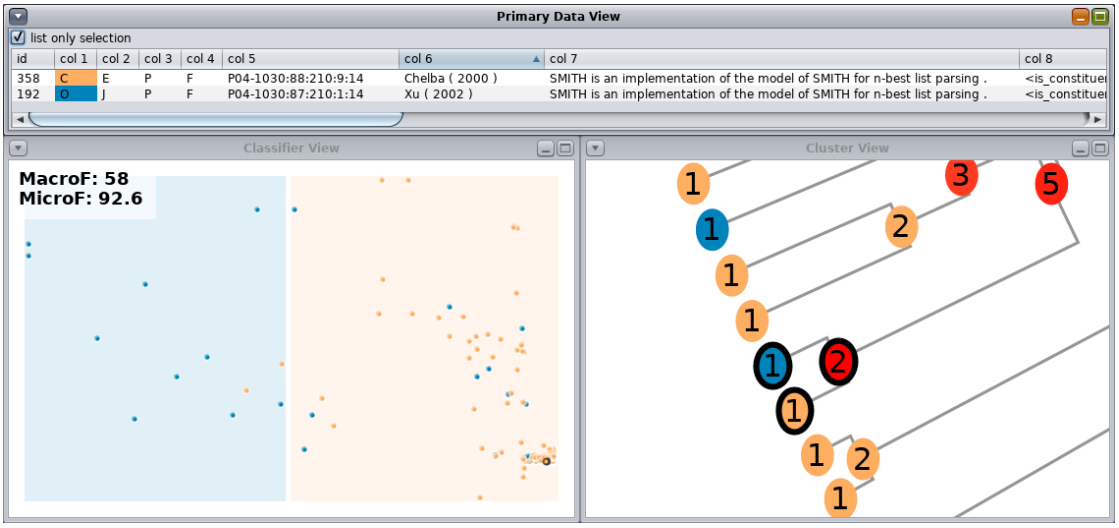


Figure 6.2 — Screen shot of the primary data, the classifier, and the cluster view with an example of two citations that occur in the same sentence. The lower one in the primary data view gets misclassified. In the primary data table, col. 5 contains a key for the citing paper, col. 6 contains the citation key of the cited paper, and col. 7 contains the sentence with the citation. *SMITH* is a placeholder for the actual citation key in the sentence.

the defined features is a bag-of-words representation of the sentence enclosing the citation and most of the other features also depend on the surrounding sentence results in this problem. Figure 6.2 shows one example of two citations occurring in the same sentence. In the classifier view, both instances get positioned at the exact same spot, thus only one of them is visible and the one labeled *operational* is consequently misclassified. We find many similar examples while exploring the data set. The issue could be solved by splitting up sentences containing more than one citation as constituents. In the table of the primary data view, a parse tree of the sentence of each citation is stored. It can be used to assign the largest constituent containing the respective citation as the basis for feature extraction. Unfortunately, we are not able to test it right away with the Stanford language, but we disabled the bag-of-words feature for the rest of the session.

The next finding is a cluster of four citations, of which two are labeled *operational* and two are labeled *conceptual*. By looking at the primary data view, which contains the sentences surrounding the citations, and the citation keys, we see that one of the citations labeled *operational* refers to data sets, the other one refers to a specific type hierarchy that the citing paper adopts from the cited paper. The citations labeled *conceptual* refers to aspects of two different question answering

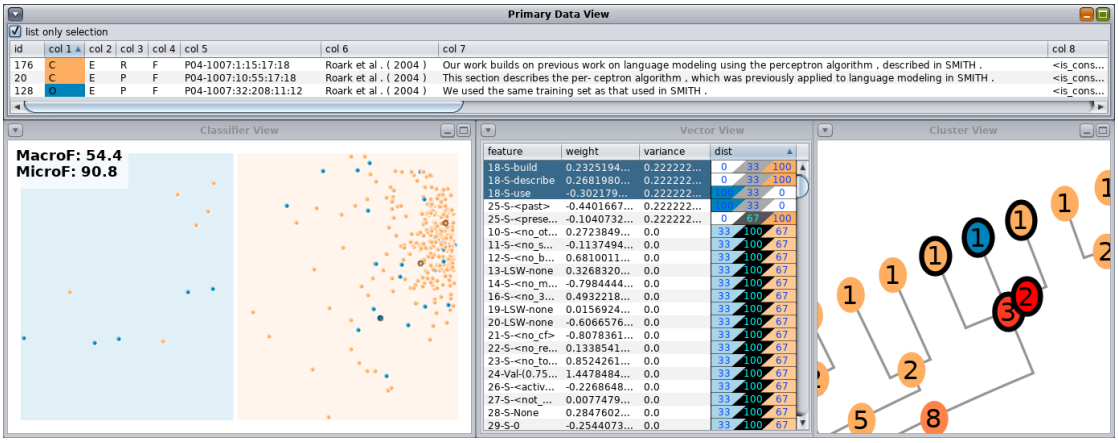


Figure 6.3 — This screen shot shows the cluster containing three citations to the same paper. The first two citations in the primary data view refer to the classification algorithm used in the cited paper, and are correctly labeled as *conceptual*. The last one refers to the training set and is correctly labeled as *operational*. In the classifier view, the three examples cannot be separated correctly, and all of them end up on the *conceptual* side. The features for the three verbs *build*, *describe*, and *use* are highlighted in the vector set view.

systems. Although all of those labels are correct, we realize that it is generally not obvious where to draw the line between a tool borrowed and a concept adopted from another work. The next finding supports the assumption that such instances are problematic for annotators. It is a cluster with two citations, one citing a paper about a classification library and the other referring to a corpus. The citation to the classification library is correctly labeled *operational*, whereas the citation referring to the corpus is erroneously labeled *conceptual*. Based on this, we plan to update the annotation guidelines with the instruction to label data sets used from other publications as *operational*, and emphasize that *conceptual* aspects of systems that are adopted by the citing paper are to be marked as *conceptual*. As examples, we can use the citations just found.

Next, we discover a node with three citations in the cluster view. One refers to a Prolog implementation and another refers to a parser. Both are labeled correctly as *operational*. The third citation in the cluster refers to a set of metrics and is labeled *conceptual*. This is also a borderline example, for which the annotation guidelines offer no clear guidance, but which should have been labeled *operational*. We will further refine the guidelines using this citation as an example.

The next cluster that attracts our attention contains three citations. Two of them are correctly labeled as *conceptual*. They refer to an algorithm used in the cited

article. The third citation refers to the training set used in the cited paper. It is labeled correctly as *operational*. In the classifier view, however, we can see that the classifier is not able to correctly separate these examples and all of them end up on the *conceptual* side of the decision boundary. Figure 6.3 shows this. Despite the fact that these particular examples are labeled correctly by the annotators, we realize that the annotation guidelines do not have clear instructions on how to label algorithms that the citing work builds on. They should be labeled as *conceptual*, and we will update the guidelines consequently. We then take a look at the main verbs of the sentences containing the citations. They are defined as features and are thus visible in the vector set view (see Figure 6.3). The two sentences that are labeled as *conceptual* have the main verbs 'describe', with the cited classification algorithm as direct object, and 'build [on]', with the classification algorithm as prepositional object. The *operational* citation has the main verb 'use', with the training set as direct object. In the vector set view (Figure 6.3), we see that the classification model had weights pointing to the *conceptual* class for the two verbs 'describe' and 'build', and a weight pointing to the *operational* one for 'use', which is what we expect the classification algorithm to learn. However, we realize that single verbs are probably not features that offer the right granularity for classification, because they run into data sparseness problems. We expect to cause the classifier to put more emphasize on these verbs by introducing verb classes as features, thus reducing data sparseness. For this, we plan to build on verb classes from the VerbNet Schuler [2006] project.

6.5 Future Directions

In this chapter, we discussed FeatureForge, an approach designed to support free exploration and sensemaking of language data for feature development and corpus improvement. It uses interactive visualization to support NLP researchers and practitioners in two aspects: (i) spot problems with the feature set and define new features to improve discriminatory power and thus classification performance, and (ii) find groups of instances that are hard to label or get systematically mislabeled by the annotators. This helps to revise the annotation guidelines and add newly found instances for better illustration of particularly hard examples. The use case based on a citation corpus shows that our approach effectively helps with those two aspects. By facilitating interactive exploration of corpora with a special focus on the labels of the instances for a specific classification task and the feature representation of the linguistic instances, users of FeatureForge are supported during the process of designing features and in addition they can easily spot annotation problems. We expect that future extensions of our approach will support the development of new and better performing machine learning based NLP methods. These can

then, in turn, help improve interactive and visual analysis approaches to languages according to the two level model (Chapter 2) of text analytics.

As described above, the FeatureForge system is a prototypical implementation of our approach, and currently has some restrictions. Future research is necessary to broaden the concept and make it more versatile for a larger range of machine learning problems in NLP. Firstly, it can be extended with a more powerful feature definition language, such as the one sketched by Kobdani et al. [2010]. Like the language of the Stanford Classifiers, it is based on primary instance data in a tabular format, but offers much more flexibility to extract and define features based on this data. Furthermore, primary data is currently static during the runtime of the system. This can be extended and made dynamic by integrating linguistic tools that can be directly used as a data source for columns of the primary data table. Here, analysis results of VA methods can also be included as data sources to support NLP experts with creating primary data based on their own insights gained about the text data.

Secondly, a broad evaluation of the prototype system would help to shed more light on how different aspects help users with their task. Since FeatureForge is tailored to a specific user group, namely NLP specialists, a quantitative study to investigate the usefulness of our approach will not be possible. The reason for this is that it is hard to find NLP tasks and corpora with which a large number of NLP specialists are equally familiar. Nevertheless, a broad qualitative study with NLP practitioners can be done to learn about the benefits and deficiencies of the FeatureForge system.

Thirdly, there has been a resurgence of neural network based learning approaches under the name of deep learning in NLP, but also many other application areas of machine learning. These methods reduce the need for feature engineering significantly as they are able to learn adequate data representations [Bengio et al., 2013]. However, at the same time, data curating and labeling becomes even more important, as these methods have a particular need for large training sets to achieve satisfactory results. In addition, these methods produce complex models. Here, visual interactive approaches can help experts produce insights into what these models have learned from the data to better understand them, and use this gained knowledge to improve them further.

CHAPTER 7

Exploration of Scientific Literature and its Context

This chapter contains previously published material from the following publications:

F. Heimerl, Q. Han, S. Koch, and T. Ertl. CiteRivers: Visual analysis of citation patterns. *Transactions on Visualization and Computer Graphics*, 22 (1):190–199, January 2016a. doi: 10.1109/TVCG.2015.2467621

The methods that we discussed so far have all been designed with scientific literature documents and data sets as their main target. However, all of the resulting visual interactive data analysis and exploration approaches have turned out to be highly versatile in that they can be applied to a broad range of text documents and data sets. This is largely due to the fact that we have specifically concentrated on the text data and natural language aspect of scientific literature analysis. As the problems that exist in this particular domain are not generally different from other text types and domains, the solutions and approaches that we developed to support free exploration and human sensemaking of textual data are applicable to a broad variety of texts. The final VA approach that we discuss as part of this thesis is different in this respect, as it tackles a problem that is specific to scientific literature

and related types of documents, such as patents. This problem is the interactive exploration of the embedding of documents into their scientific and societal contexts. With the approach presented in this chapter, we aim at supporting human users with exploring contents and trendy topics of scientific literature data sets. This information is visually linked with the scientific surroundings of these publications based on their citations. We specifically concentrate on document meta data, including citations links, to provide information about a document's context. This helps users to learn about the important research questions of a field over time, the main authors contributing to certain topics in a field, and its connections to other communities and field based on the citations.

7.1 Motivation

Awareness of thematic and structural changes and developments in a scientific community is important to create visions for its future. This enables researchers to develop new research questions and ideas that keep a field thriving. Understanding those dynamics is important for people new to a research field just as it is for long term members. The former can identify key topics, authors, and publications to gain inspiration for their own work, and learn how to position their publications. The latter are interested in the influence and importance of past developments on current trends to better reflect on and gauge the future evolution of the field. In order to gain new insight into a scientific community, users have to base their inquiry on its past and current scientific output in the form of publications. For a comprehensive picture of a field, users require flexible access to various aspects of these publications and interactive visual support to detect patterns. In particular, knowledge about thematic dynamics and their interactions with other scientific disciplines are crucial factors that enable researchers to have long term success in a field.

Approaches to the interactive visual analysis of scientific literature that were published prior to ours have ignored this link and consequently provided an incomplete picture of a discipline. In particular, they primarily focused on visual analysis of citation networks, their development, topic or thematic evolution, and the creation of science maps for overview. Our approach, in contrast, considers the outreach of research to other communities relative to the topic dynamics of a field. We achieve this by taking into account sets of venues that are co-cited from within an area of research. Sets of venues have the advantage of being directly interpretable and they can be derived from an available set of publications, without requiring to get access to and deal with additional, large document repositories.

Great value is added if the mentioned outreach can be tracked back to themes and topics developed in a scientific field over time. Previous interactive analysis techniques proposed for scientific documents have concentrated on one of two aspects: document contents and topic structure, sometimes paired with metadata analysis, and citation networks, typically depicted as node-link diagrams. While these approaches are effective for analyzing different facets of the data, they do not support correlating content of publications with their citations. They thus miss an important aspect to support sensemaking in the scientific literature domain and help users derive comprehensive knowledge of a field.

We address these shortcomings by linking topics and citations, and facilitating their combined exploration with adjustable automatic methods to extract thematic content and citation patterns over time. For this, we describe CiteRivers, a VA approach that supports users with sensemaking and free exploration of publication sets by helping them to better understand the dynamics within a scientific field. It is based on a new technique for the visual combination of the contents of publications with their citations. We integrate the popular visual streamgraph metaphor to depict thematic developments over time, and augment it with visual links to cited venues. Through interaction with the visual abstractions of the data set, users can correlate and filter different aspects of the documents. This allows them to iteratively drill down to different aspects of the data set, thus supporting sensemaking processes. The approach further comprises automatic methods to join, correlate, and aggregate data from multiple sources. In addition, users are given full control over the granularity of the automatically created abstractions to adapt them to for their goals and interests.

7.2 Related Work

Most of the work related to this approach has already been discussed in Chapter 2. For this reason, we only mention the most closely related publications, and sketched the differences to our approach. The InfoVis contest [Plaisant et al., 2008] held in 2004 has produced some approaches related to ours. Ke et al. [2004] focus on the exploration of citation networks and the interrelationships of influential authors, but do not combine it with text analysis. Wong et al. [2004] apply IN-SPIRE to the InfoVis data set to track research topics over time, but do not include citation analysis. PaperLens [Lee et al., 2005] also allows to track research topics and their popularity over time and shows the most often cited authors and papers per year. These three papers are related to our work as they include methods to explore and track topics over time, and identify the most prolific authors from a set of publications. Compared to the three systems, CiteRivers provides a new abstract representation of citations based on community structure that is correlated with

the topic dynamics of the analyzed document set. Our approach also features a visualization of the publications per topic for the most prolific authors. Ahmed et al. [2004] use a 3D representation of topics over time and depict the citation links between them as straight lines. It differs from our approach in that it does not aggregate citations and only handles the ones contained in the data set at hand. CiteSpace II [Chen, 2006], which is entirely based on citation data, supports the analysis of research fronts and their intellectual bases. The latter describe the set of publications that are fundamental to a scientific field, while the former comprise cutting edge research. The Eigenfactor project [West, 2010] uses spectral methods for citation network analysis to identify important journals across disciplines and analyze their intra- and inter-discipline citation links, but it does not take the publication contents into account. Dunne et al. [2012] support scientific literature search and the exploration of major topics in a research field. Görg et al. [2013] present an approach to visually and interactively correlate documents and entities based on meta data and content. Other projects just focus on meta data, such as citations [Zhang et al., 2009; Stasko et al., 2013] or comprehensive bibliographic information [Dörk et al., 2012; Beck et al., 2016].

CiteRivers features an extended version of a streamgraph to depict clusters over time. They have been originally introduced by Havre et al. [2002] as a visual metaphor to convey thematic developments in large document collections. They have since been applied to various other types of data, such as baby name popularity [Wattenberg, 2005], and have even found their way into the mainstream media [Byron and Wattenberg, 2008]. Extended versions show thematic changes and interaction between topics in text data sets [Cui et al., 2011], and Wu et al. [2014] combine them with sentiment analysis results encoded by color. We extend streamgraphs with a flowgraph-based exploration technique. Flowgraphs have a long history in infographics, e.g., as a map overlay to visualize the movement of goods or people. They have recently become popular in combination with new interaction methods for large tabular displays [Tobiasz et al., 2009]. Phan et al. [2005] discuss a flowgraph creation method that uses hierarchical clustering in a way similar to ours.

7.3 Visual Analytics of Topics and Citation Patterns

CiteRivers, our approach for the exploration and citation patterns consists of five views of the document set. Its implementation is depicted Figure 7.1. All views are connected by brushing and linking to allow for an easy combination of the different data aspects. The two central views, that implement our approach to link

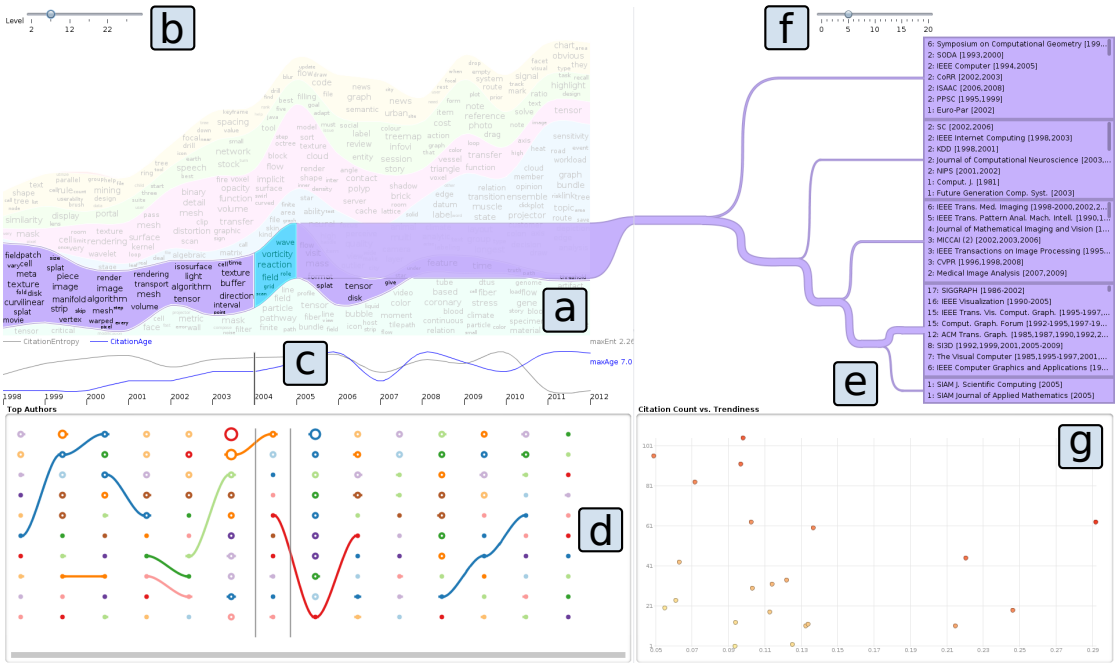


Figure 7.1 — The desktop of CiteRivers consists of the stream panel (a), the level slider (b), the citation aggregation panel (c), the author panel (d), the citation flow panel (e), the categories slider (f), the document trend plot (g).

document contents and cited communities, is the *streamgraph panel* (Figure 7.1 (a)) and the *citation flow panel* (Figure 7.1 (e)). The former is situated in the left upper space of the desktop and contains groups of the publications depicted as streams of varying prominence along a time axis. The latter is located to its right and shows an aggregation of the citations of the documents in a selected time step of a stream. All views on the left side of the desktop (Figure 7.1 (a)/(c)/(d)) are aligned with the time axis as they contain time-dependent information. The views on the right side (Figure 7.1 (e)/(g)) display additional information for each focused time step of a stream.

7.3.1 Streamgraph and Citation Flow

The *streamgraph panel* (Figure 7.1 (a)) visualizes the topic structure of the data set with the popular streamgraph method [Havre et al., 2002]. It is an aesthetically pleasing and easily interpretable visualization scheme for multiple time series, conveying individual values as well as their sums [Byron and Wattenberg, 2008]. In addition, the flow metaphor of streamgraphs makes it possible to combine them

smoothly with a flowgraph, which we integrated in order to link topics and cited communities, without breaking the metaphor. This results in a straightforward, organic visualization that users can interact with naturally and smoothly. In it, users can explore thematic clusters along the time line. CiteRiver supports two different ways of grouping publications that can be selected according to the user's wishes. One clusters documents hierarchically based on content using the spectral clustering scheme described in Chapter 2. The second method for grouping uses metadata attributes of the documents. It allows users to group documents, for example, by the conferences they were published at, or the affiliations of their authors.

If the user chooses the hierarchical method, the level in the clustering tree can be switched interactively with the *level slider* (Figure 7.1 (b)). This changes the number of clusters to a granularity that suits the user based on her current exploration goals. In case larger time spans are selected, the binning of documents can be made coarser than one year to maintain visual scalability. Users can mouse over and explore the different clusters in the streamgraph. When mousing over a specific stream, it becomes focused and is highlighted through higher color saturation (see the purple stream in Figure 7.1) compared to the other streams. This updates the views on the left side of CiteRiver's desktop to show the information of the highlighted documents in the stream. The areas for each of the time slices are separate visual elements that users can interact with, e.g., the turquoise area in Figure 7.1. We call these the blocks of the streamgraph.

Each block contains a word cloud that gives an impression of the contents of the publications that it contains. The terms are extracted from the abstracts of the publications using the G^2 measure discussed in Chapter 2. We place the terms along a spiral path, starting with the most frequent term at the center of the block [Viegas et al., 2009]. This results in visually salient labels at the center of the block, and in visually appealing word clouds. In case the blocks are so small that only few terms fit into it, users can mouse over a block to highlight it. This triggers a tooltip with a larger word cloud in a line by line layout, as depicted in Figure 7.5. When a block is highlighted, it is marked with a turquoise background, and extended and attached on its right side to the *citation flow panel* (Figure 7.1 (e)). The left border of the focused block extends to the time line axis below the streamgraph panel, where it marks the corresponding year.

The *citation flow panel* (Figure 7.1 (e)) is the second component of our approach. It is attached to the right of the streamgraph panel and displays an aggregation of the references of a selected block's documents. The aggregated citations are visually linked to the block, maintaining the stream metaphor. We like to see the citation streams as rivers that flow from the cited conferences or journals, i.e., the leaves of the flowgraph, to the streams of the streamgraph, with small tributaries

joining to form a larger river that swells up to the root node of the clustering and finally ends within the block of the focused stream. We have decided to use streamgraphs, as they are effective to convey movement of objects or material from one point to another while keeping clutter at a minimum [Phan et al., 2005]. To further reduce clutter and increase readability, our layout algorithm introduces only binary splits. The result is easy to interpret and follow, and thus helps users to understand the distribution of citations from a focused block. When a block is focused, the extension to its right is the end of a flow out of the citation flow panel. This is depicted in Figure 7.1 for the highlighted block. The flow depicts knowledge from the cited communities that flows towards the citing documents, with individual streams coalescing on their way to the selected block. In the citation aggregation panel, multiple smaller flows start at clusters of publication venues. These contain similar conferences and journals listed by their names and the years of the citations.

The entries in each list are sorted according to citation count, indicated by the number preceding the venue name. Details about each venue can be accessed by double clicking on a list entry. This opens a browser with the DBLP [Ley, 2002] page of the conference or journal. It offers ample information, including all authors and publications for each installment or issue, and links to the conference or journal web page. Users can select single venues or entire clusters, showing citation numbers for each block of the streamgraph (this is depicted in Figure 7.6). The split-up of the stream is based on a clustering of all conferences and journals from the DBLP database according to community structure. Again, users can interactively adapt the granularity of the communities by setting their number with the *categories slider* (Figure 7.1 (f)). For example, on a low granularity level, visualization venues (e.g., IEEE InfoVIS or IEEE VIS) are in the same clusters as venues primarily focused on rendering (e.g., Eurographics or SIGGRAPH), but they split when granularity increases.

Changing granularity causes the tree to be laid out again, but with the clusters remaining in the same order and approximately retaining their positions. The vertical space is distributed to the lists relative to their element numbers, with scrollbars if some elements are hidden. In addition, the publications citing the selected venues are highlighted in the document trend plot (Figure 7.1 (g)). The community clusters are connected to the selected block of the streamgraph through a binary cluster tree that is laid out as a flowgraph. Its edges are drawn in a rounded fashion with their thickness scaled according to the number of the cited venues they dominate. The layout algorithm assigns equal horizontal space to both children of a node, starting from the root node. All available space is thus evenly distributed, which results in a balanced layout.

The *citation aggregation panel* (Figure 7.1 (c)) is located right below the streamgraph panel. While the citation flow panel visualizes local distribution of citations per block of the streamgraph, users also need an aggregated view of the citation behavior over time. This helps them get a quick overview during exploration and provides them with information of what to explore in more detail. The citation aggregation panel contains plots of two characteristic values of the citations in each block along a highlighted stream. We have chosen these measures to help users track the evolution of citation behavior and point them to potentially interesting blocks. The blue curve shows average citation age of each block of a stream. It is the average age of references of the publications of a block at the time the document was published. From this curve, users can learn how far back publications in a specific year and stream are reaching with their citations. It might also indicate how new and trendy a topic of a stream is, as less trendy topics will be based on work that is potentially older, while documents addressing trendy topics will more likely cite similar recent publications. The second, gray, curve depicts the citation entropy along the highlighted document stream. It is calculated based on the different conferences and journals cited in the focused block according to Equation 7.1, where V is the set of cited venues, and $\#cites(v)$ denotes the number of citations to venue v in the block. The citation entropy measures the diversity of the citations of the publications along the stream.

$$H(V) = -\sum_{v \in V} \frac{\#cites(v)}{\sum_{v' \in V} \#cites(v')} \log \frac{\#cites(v)}{\sum_{v' \in V} \#cites(v')} \tag{7.1}$$

This helps users assess how widespread documents in a stream cite publications from different academic disciplines and how this changes over time. The entropy, for example, rises if documents in a stream start to cite publications from a new scientific community in addition to their traditionally cited fields. To make orientation easier for users, the citation aggregation panel features a vertical line that marks the year of the currently highlighted block in the streamgraph.

7.3.2 Author Panel

The *author panel*, shown in Figure 7.1 (d), is situated right below the citation aggregation panel. While we were discussing and designing our approach, it became evident that users who know a community typically connect research topics with authors. Our approach consequently includes a view that lets users refer to the most prolific authors of each block of a stream. We show the ten most highly ranked authors of each block to give users orientation and support them with understanding and interpreting topic streams. While this number could be increased, we found that ten is a good number in practice that provides ample references to authors, but does not overwhelm users with too many choices. We measure author prolificacy

by the number of publications authored per year. In the author panel, circles of different sizes are arranged into a matrix, with each column corresponding to one block. Each column contains up to ten circles that represent authors, ordered according to their rating. The size of each circle represents author's rating. When a circle is moused over, it displays the name of the corresponding author right above the circle. To color the circles, we relied on ColorBrewer2 [Harrower and Brewer, 2003] for a color mapping of twelve distinct colors. This means that we have to reuse colors for different authors. In order to avoid confusion, we link authors that occur in a sequence of adjacent years with a curve of their respective color. If authors do not occur within the matrix in adjacent years, but further away, we add a stub to the left or right of their circle. Similar stubs are used by [Collins et al., 2009c] to indicate links between terms in their Parallel Tag Clouds. A stub indicates an outgoing edge to an earlier or later year. Stubs are connected when the circle is moused over, linking all instances of the corresponding author. Thus, users can easily track a specific author over time. Authors can also be selected by clicking their circle. All blocks in which the selected author has published are then highlighted in the streamgraph panel and show the respective number of publications. This is depicted in Figure 7.8. In addition, all publications of the selected author are highlighted in the document trend plot from Figure 7.1 (g).

7.3.3 Document Trend Plot

The *document trend plot*, depicted in Figure 7.1 (g) is situated on the right side of the desktop, below the citation flow panel. When exploring the topic distribution and dynamics of a data set, users need access to the single documents. This is important for sensemaking processes because it lets users gain new insights by directly referring to abstracts or full documents. It also supports them with finding new work that matches their research interests. To give users orientation in the document set backing a block of the streamgraph, we show a document scatter plot that layouts the documents according to two dimensions. These are citation count of each documents and a trendiness score, which provide information about the popularity and success of each publication. While the first one is a quantification of popularity, the latter quantifies the novelty and success of ideas within a paper, as described later. This score can be seen as the local extension to the global topic dynamics in the streamgraph. Rarely cited publications with a low trend value reside in the lower left corner of the plot, while highly cited ones with a high trendiness score populate its upper right corner. Thus documents with less trendy content are placed on the left side of the plot, towards its past, while more trendy ones are placed closer to their future towards the right edge of the scatter plot.

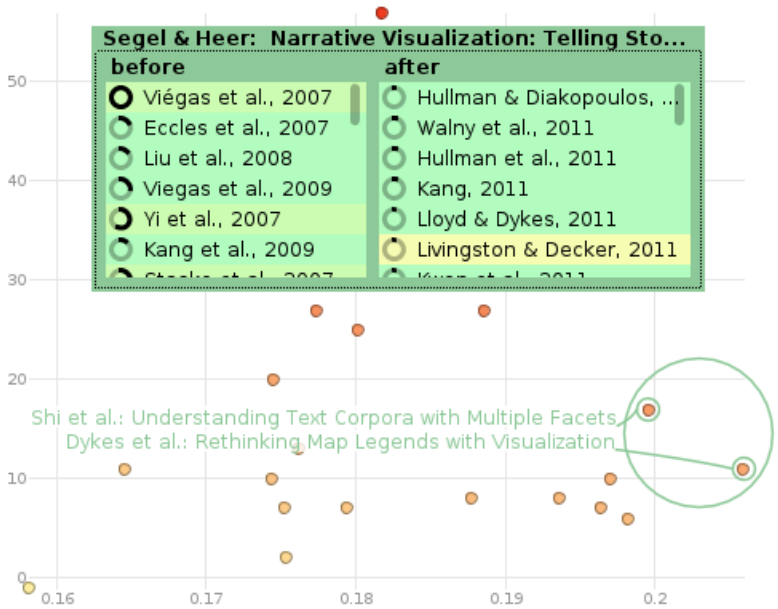


Figure 7.2 — The document trend plot with the lens showing authors and title for two documents. The detail panel has been activated for the publication right above the panel. It lists its most similar documents.

Users can explore the space by using the radial title lens, as can be seen in Figure 7.2. The lens can be activated by clicking into the free space of the plot. It shows authors and titles, visually linking them to the respective document. The lens has some similarities with the technique discussed in Chapter 4, but is not used to characterize multiple documents in the scatter plot. We rather included it to avoid clutter by showing all document titles at once. It can thus be considered an excentric labeling technique [Fekete and Plaisant, 1999]. The title lens always has the same color as the currently highlighted stream, and its size can be adapted using the mouse wheel. To learn more about the depicted document set, users can click on the document glyphs to activate a detail panel, which is also shown in Figure 7.2. Apart from author and title of the document, these panels show two lists of the documents that are most similar to the selected one. The left list contains those published earlier, while the right one contains those published later. As the trendiness score for the documents is computed based on them, the detail panels serve as an explanation for the scores. The listed documents are colored according to the stream they are part of, which gives users information about the distribution of the topics in this set of documents. Relative citation counts for the listed documents are depicted as radial donut charts, normalized to the largest count in the panel. Absolute counts are available via tool tips for each entry.

7.4 Data Processing

The presented approach includes data from multiple sources that is joined and aggregated. This section presents and discusses the different data sets and the processing and aggregation steps. The chosen techniques are known to work well for extracting and comparing textual content. Figure 7.3 contains a diagram of the entire process.

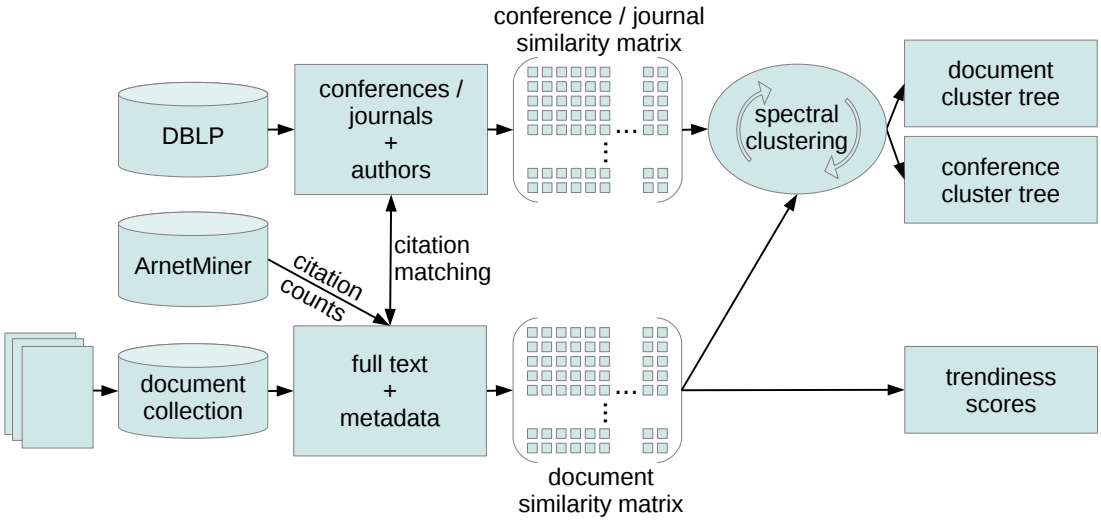


Figure 7.3 — CiteRiver's data processing pipeline with three data sources: (1) the raw publications for full texts and metadata, (2) the DBLP database for publication venues, and (3) the ArnetMiner database for citation counts. From those, we create similarity matrices of documents and venues to feed them to the spectral clustering algorithm and compute the trendiness scores.

7.4.1 Publications and Trendiness

For an exploration session with CiteRiver, users select a document set that they want to explore. This collection of scientific publications is either available in a structured format, or has to be extracted from pdf files or even scans. Our approach is applicable to all data sets that contain the full texts of the documents, metadata including authors, titles, and abstracts, and complete reference lists. As shown in Figure 7.3, each of the documents in this data set is matched against the ArnetMiner [Tang et al., 2008] database to extract citation counts. The ArnetMiner database comprises all of the entries in DBLP enriched with additional information, including citation counts. In the next step, the text content of the documents is transformed into the vector space model (see Chapter 2).

Before we can create the vectors, we linguistically preprocess the texts by applying a tokenizer, a lemmatizer, and a stop word removal scheme. Tokenization separates single tokens from the sequence of characters that represents the text. Then, a lemmatizer transforms these tokens into their lemmas, the base or dictionary form of a word. This step removes, e.g., plural forms of nouns, and conjugations of verbs. For both tokenization and lemmatization we use the Stanford CoreNLP package [Manning et al., 2014]. The next step is a quite aggressive stop word removal method. Instead of traditional stop word removal, which uses a fixed list of words that contain no information in isolation, we remove all terms that occur in more than 60% of the documents. We found that this removes many frequently used words that introduce noise and results in more compact and informative vectors that help to distinguish better between documents of different topics. Based on the resulting vectors, we compute a matrix of pairwise document similarities using cosine similarity (see Chapter 2). The resulting matrix serves as the basis for the document trendiness score, and can be fed into the clustering algorithm to create the thematic clustering of the data set (see Figure 7.3). If the user wishes, the clustering algorithm can be disabled, and the documents can alternatively be grouped according to various metadata aspects, as described previously.

Trendiness Score

The trendiness score gives an impression of the freshness of ideas in a publication and their dissemination into later ones. It thus captures one important aspect of publication success that is independent of citation numbers. Our score is based on document similarity and is thus akin to Shaparenko et al.'s lead/lag index [Shaparenko et al., 2005]. It is, however, more flexible because it considers all available past and future documents, while the lead/lag index relies on a fixed neighborhood of size k in a document's past and future. Another measure for publication success is presented by Chen et al. [2007]. It is an adaption of the h-index [Hirsch, 2005] and is thus entirely based on citations. As the document trend plot (Figure 7.1 (g)) already combines raw citation counts with the trendiness score, such a citation-based measure would not introduce new information into the visualization. Combining citation counts and trendiness gives users a deeper insight into the context of a publication than each measure individually. Take literature surveys as an example. They tend to attract numerous citations as they summarize the state-of-the-art in a field, but typically do not contain any technical innovations. Literature surveys thus get high citation counts with low trendiness scores. Their impact consequently cannot be captured adequately by the individual measures alone.

To quantify the freshness of ideas of a publication d_i, we take a look at the set of documents published earlier, $D_{before} = \{d_x \mid \text{earlier}(d_x, d_i)\}$. We estimate the

influence of earlier documents based on their similarity to d_i and calculate the impact score I_{D_{before},d_i} of D_{before} on d_i according to Equation 7.2.

$$I_{D,d_i} = \sum_{d_j \epsilon D} e^{-\frac{s(d_i,d_j)^{-2}}{\tau^2}} \tag{7.2}$$

Depending on whether the documents in D are older or newer than d_i, Equation 7.2 quantifies the influence of D on d_i, or of d_i on D, respectively. The influence measure is based on the cosine similarity of documents $s(d_i, d_j)$, assuming that influenced publications imitate the influencing ones to a certain extent. Though not capable of capturing all aspects of scientific interaction within a community, this assumption is able to measure a publication's impact without considering citation count. After estimating how much the ideas in d_i become popular in future publications, $D_{after} = \{d_x \mid \text{later}(d_x, d_i)\}$, by computing I_{D_{after},d_i}, the overall trendiness score is determined by weighting the influence on d_i against the influence of d_i on future work (Equation 7.3).

$$\text{trendiness}(d_i) = \frac{I_{D_{after},d_i}}{I_{D_{before},d_i}} \tag{7.3}$$

The Gaussian kernel, whose size is defined by τ, determines the similarity level at which we assume that one publication influences the other. We experimentally determined a value for τ by iteratively refining it on our data set, resulting in $\tau = 10$.

7.4.2 References and Communities

In our approach, we aggregate citations by grouping publication venues by scientific communities. Each reference of the publications in the data set is mapped to the group that contains the venue it was published at. Previous work exists that extract scientific communities based on citations, co-authorship, or content, for example by Newman [2004] and Chen [2006]. Thus our approach to detect the influence of different research disciplines based on cited venues and author overlap between them, has been novel at the time we first published it. And to the best of our knowledge there is still no comparable approach.

Hierarchical Communities

We extract communities based on the DBLP data set [Ley, 2002] of computer science publications. It is one of the largest available data sets of such publications. In addition to publications, DBLP has entries for publication venues, linking each of the documents to the conference or journal it was published at. As depicted in

Figure 7.3, we extract all conferences and journals contained in DBLP. In addition, we collect the author names of each document and the associated venue. We keep each issue of a journal or installment of a conference as separate entities, allowing us to better model the thematic dynamics of venues over time.

In the following step, we create a similarity matrix for the extracted venues based on their author overlap. This is a way of modeling community affiliations of conferences and journals. We assume that the more authors who publish at both venues, A and B, the more similar both venues are in terms of the scientific communities they are part of. We model this using the Jaccard coefficient as a similarity measure for venues. The Jaccard coefficient is a way of quantifying the overlap between two sets of entities: $jaccard(A, B) = (|A \cap B|)/(|A \cup B|)$. Spectral clustering is applied to the resulting matrix to create a hierarchical community structure.

We chose to use conferences and journals as the base categories to create a community hierarchy for two reasons. Firstly, we have decided against the obvious alternative of creating and clustering a co-author network from DBLP (such as, e.g., Newman [2004]) because DBLP lists about 1.5 million authors. Partitioning the resulting large co-author graph would take enormous computational resources and thus processing time. In addition, we would only use the first few levels of the resulting hierarchy, with the rest being too fine grained and therefore irrelevant for us. Another disadvantage of co-author networks is that we would have to tackle the hard problem of author name disambiguation [Ferreira et al., 2012], as two authors with coincidentally identical names would distort the partition results. Using larger entities, such as publication venues, as a basis is much more robust against this problem, and we found that we can get sufficient results by just treating these ambiguous names as noise. The second reason we have decided to use conferences and journals as the base categories of our community clustering is that their names are much easier for users to interpret compared to the names of single authors.

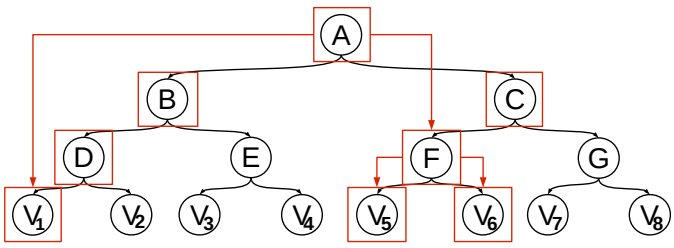


Figure 7.4 — Community tree for eight conferences (V_1 to V_8) with an overlay for a block that contains V_1, V_5, and V_6. The marked nodes are dominating a relevant venue. The red overlay tree includes all marked nodes that have either two or no marked children.

Community hierarchies are stored, and used to aggregate citations each time a user highlights a block. The result is then visualized in the citation flow panel (Figure 7.1 (e)), showing only the venues of references of the highlighted block. The aggregation is created by computing an on-the-fly overlay on the community tree containing only venues relevant for the block. As depicted in Figure 7.4, this is done by recursively marking all clusters in the tree that contain at least one of the relevant publication venues. In a second step, a new, temporary tree is created, spanning only those nodes that contain relevant venues and either have no or two children. The resulting partial community hierarchy is then depicted in the citation flow panel (e) of Figure 7.1. As the clustering for each set of references is always based on the same basic tree, contradictory combinations of venues in two different blocks cannot occur. This helps users with sensemaking by allowing them to establish a consistent mental map of the community structure of publication venues.

Reference Extraction

To map the references of each of the publications in the data set to its community cluster, we extract the conference or journal of each referenced document. Using the venues from the reference strings is not possible, because venues are not referenced in a standard way, and authors use different names and abbreviations for the same venue. To solve this problem, we use the DBLP database again, and find the corresponding entry for each reference. These entries are linked within DBLP to the conference or journal they were published at, giving us unambiguous identifiers for each venue. To find the entry for a reference, we use its title string and compare it to all DBLP entries. For this, we use the Levenshtein distance [Levenshtein, 1966] that measures string distance as the number of character insert, delete, and exchange operations needed to convert one string into the other. A title is matched to its most similar DBLP title above a threshold that we determined iteratively by manually reviewing the results. This fuzzy matching mechanism allows us to handle small differences in the strings caused, e.g., by orthographic variations. With this technique we were able to match approximately 70% of references to DBLP entries. All unmatched references are currently ignored by our implementation. Once we have mapped a reference to its associated venue in DBLP, we can use this information to create community trees for each block with the method described above.

7.5 Use Case

This section presents an example exploration sessions that showcases the capabilities of our approach and its implementation. We first describe the preparation of a data set we created for this use case, and then describe an example analysis session step by step. Although the use case includes all of the features and elements of the approach and our prototype, it focuses on the benefits of linking topic dynamics and aggregated citations, as this is the main contribution of this work.

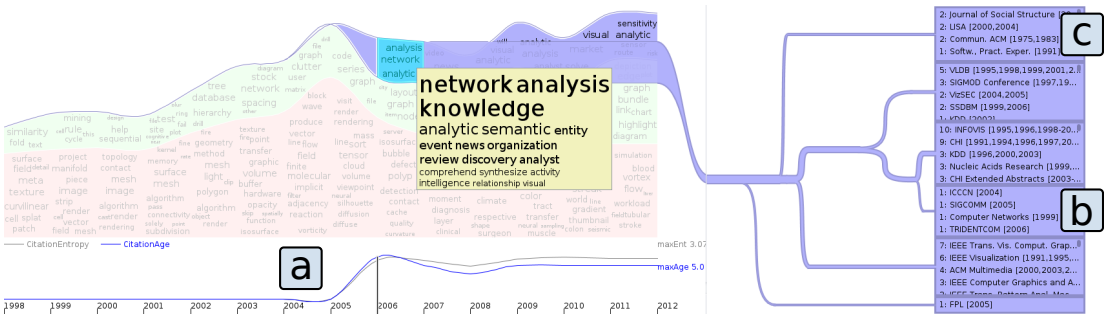


Figure 7.5 — For the first year of VAST, the citation aggregation panel (a) shows a dent for both average citation age and entropy (2007–2008). The tool tip contains a larger word cloud with further terms for the selected block. Citations to computer network venues (b) and the *Journal of Social Structure* (c) are shown in the citation flow panel.

7.5.1 Dataset

We have prepared a data set of publications from the IEEE VIS / VisWeek conferences that cover the years 1998 to 2011. It includes all full papers for the three main conferences, IEEE SciVis, IEEE InfoVis, and IEEE VAST since 2006. The set of 1336 documents, only available in pdf format, contains 390 InfoVis publications, 797 VIS publications, and 149 VAST publications. All pdfs have been converted to plain text using a commercial OCR solution[1]. We use OCR rather then extracting text directly from the pdfs, because we found that it significantly improves the quality of the resulting documents. Available text extraction tools typically exhibit problems with identifying and extracting whitespace, special characters, and the general structure of documents, especially from a two column paper format. We then used ParsCit [Councill et al., 2008] to extract the metadata from these documents, including their reference lists. The references have been mapped to their respective DBLP database entries, as described above, to extract

[1] Nuance's Omnipage Professional 18

corresponding conferences and journals. Other metadata, comprising document titles, abstracts, and author names have been checked manually, and corrected if necessary to eliminate OCR errors and assure high data quality. In addition, citation counts for each of the publications have been extracted from the ArnetMiner database, as described in Section 7.4.1. For the analysis of the data set, we have created one metadata grouping and one hierarchical cluster tree for the publications. Both can be displayed in the stream panel (Figure 7.1 (a)). The metadata grouping is based on the conference attribute of the documents. We group them by the conference they were published at: VIS, InfoVis, or VAST. This grouping gives insight into the development of the three conferences with respect to their topics over time, their interactions in terms of topics and authors, and the dynamics in citation behavior. To cluster publications according to their contents, we have used spectral clustering (see Chapter 2). The resulting cluster tree contains 50 levels, which we found sufficient for a comprehensive analysis.

7.5.2 The First Years of VAST

We start our analysis of the VIS data set by inspecting the streamgraph for the three conferences, as depicted in Figure 7.5. It shows a red SciVis stream at the bottom, a green InfoVis stream in the middle, and a purple VAST stream on top. The stream for VAST starts in 2006, its first year. A conspicuous fact shown in the streamgraph is the varying number of papers for the conferences. This is encoded by the height of the streams at each time step. We can see that acceptances for SciVis peak in 2004, and then drop again to their past level. InfoVis acceptance numbers, on the other hand, slowly rise over the years. This trend is also reflected by the official acceptance numbers for the conference, according to which it has doubled between 1998 and 2010. The acceptance number of VAST has remained steady during its first five years. We are particularly interested in VAST, the youngest

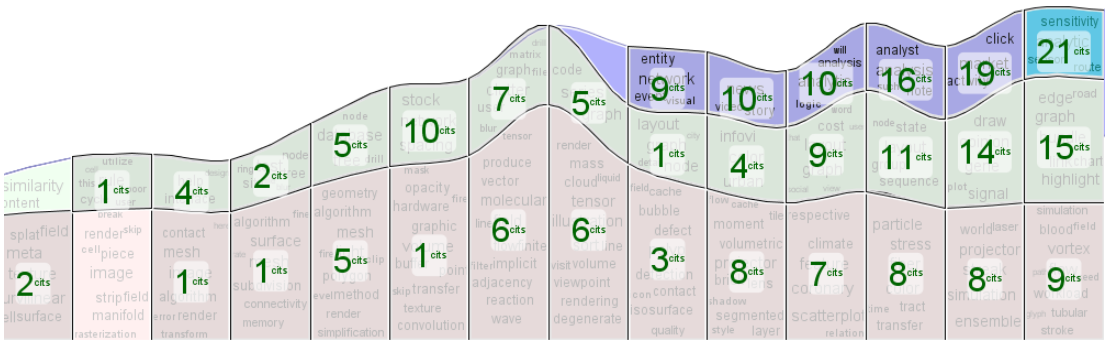


Figure 7.6 — VAST (top) has the highest number of citations to data mining venues, but citations from SciVis and InfoVis are also rising.

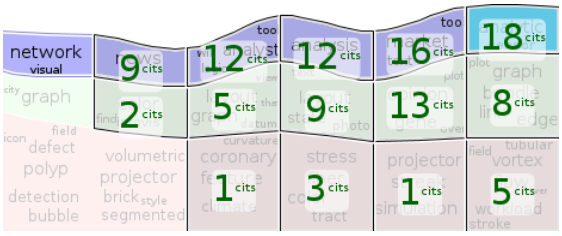


Figure 7.7 — Rising numbers of citations to VAST from VAST itself (top), but also from InfoVis (middle) and SciVis (bottom).

conference in the VIS family, whose topic is VA. Interested in the communities that VAST publications cite, we start exploring in the year 2011 using the citation flow panel (Figure 7.1 (e)). We can see that the largest cluster contains citations to expected venues from the VA and IV communities, e.g., *VAST*, *InfoVIS*, and *Information Visualization Journal*. Interestingly, there are also a significant number of data mining venues, e.g., *KDD (Knowledge Discovery in Databases)*, *Journal of Machine Learning Research*, and other data mining oriented conferences such as *Bioinformatics*, and *ACL (Annual Meeting of the Association for Computational Linguistics)* in the same cluster. These attract our attention, and to analyze them further, we separate them into their own cluster by slightly increasing granularity (from five to six clusters) using the categories slider (Figure 7.1 (f)). Selecting this cluster gives us an overview of the citations to these conferences in the data set, as shown in Figure 7.6.

Although the data mining community has been cited before by InfoVis and SciVis, VAST has a comparably high absolute number of these citations. The difference is even larger when considering relative numbers, given the fact that the number of VAST publications is lower compared to InfoVis and SciVis, from 2006 throughout 2011. We further notice that the citation numbers to data mining venues from InfoVis and SciVis also seem to increase in later years, and wonder whether this is caused by mutual influence between the nascent VA community and InfoVis and SciVis. An effect of this exchange might be the growing number of citations to VAST from InfoVis and SciVis, as shown in Figure 7.7. We discover a second hint by exploring the authors of VAST and their publication paths with the author panel (Figure 7.1 (d)). It shows the publications of a selected author for each block as an overlay on the stream panel (Figure 7.1 (a)). Figure 7.8 depicts a typical publication path of a scientist in the VIS community that starts with SciVis or InfoVis, and eventually also includes VAST publications. By exploring further authors, we learn that many of the highly prolific authors of VAST have made a similar transition. We view this as an indication of an avid exchange of ideas between the VIS conferences. During our analysis session, we discover

another interesting finding about the citing behavior within VAST. As can be seen in Figure 7.5 (a), the average age of the cited publications in its first year is higher, with a dent visible in 2007 and 2008 followed by a stable level until 2011. Comparing this with the citation numbers from Figure 7.7 suggests that the first value is higher because authors, not being able to resort to related work from previous years of the same conference, cite older material they find relevant for their work. Then, having access to fresh material from the same conference, prefer to cite these very young publications. The reason for the slight increase of the average age of within-conference cites is that the average age of the cited material grows, as the first VAST publications get older. We can see that after 2008 the average citation age levels out and stays roughly constant until 2011.

Focusing on citations so far, we now shift our attention to the thematic dynamics of the new field. We start with the word clouds that give us an impression of the topics specific for each installment between 2006 and 2011. In the first year of VAST, the term *network* is quite prominent and attracts our attention. The citation flow panel provides background information, as it includes one clusters that exclusively contains computer network conferences (Figure 7.5 (b)). In addition, the *Journal of Social Structure* in a different cluster (Figure 7.5 (c)) catches our attention. Based on these findings, we hypothesize that an unusual number of computer network analysis and social network analysis publications cause the prominence of the term *network*. Highlighting the publications that cite the respective conferences in the document trend plot (Figure 7.1 (g)) reveals four papers citing network conferences, and one citing the *Journal of Social Structure* twice. One of the computer network publications has the highest trendiness score in that year, while the other three have mid to low values. A high trendiness score indicates follow-up work influenced by the paper in later years. Selecting the cluster of network conferences shows the blocks in the streamgraph that contain publications that cite them. We see that VAST has further citations to these conferences in 2010, and we select the corresponding block for further analysis. In the updated citation flow panel, the network community cluster has grown with additional venues. We select the whole cluster, and six publications are highlighted for further investigation. Along the same lines we can follow and analyze the development and evolution of other topics, such as financial analysis and text analysis.

7.6 User Feedback and Discussion

The discussion of the effectiveness of the approach is based on feedback from six experts, three active members of the natural language processing (NLP) community, and three active members of the visualization community. We prepared a different data set for each group. The VIS data set for the latter, and a data set based on

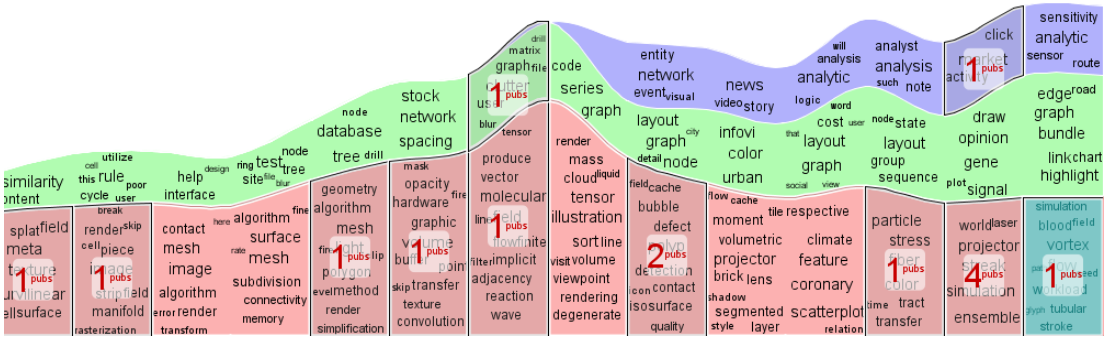


Figure 7.8 — Typical publication path of an author moving from SciVis and InfoVis to VAST.

the ACL anthology[2] for the former, using the same techniques and steps as for the VIS data set. The NLP data set includes publications of three major venues from the NLP community: the ACL conference, the *Journal for Computational Linguistics*, and the EMNLP conference from 2000 to 2013. The visualization experts were associated with our department, one with long-term experience in the field, one postdoc, and one PhD student. The group of NLP experts consisted of two postdoctoral researchers and one senior member of the field. After a tutorial of the prototype, we invited each expert to start analyzing the data set from their community, exploring whatever they find interesting with both the content clustering and the conference grouping. They voiced all of their thoughts and findings which we, based on the think-aloud scheme, recorded on paper. Finally, we asked for their opinions on possible application areas, about the interpretability of the visual representations, and the interactions. We also asked for feedback on any missing aspect of the document data that would have supported their data explorations.

All experts appreciated the CiteRiver approach as an effective top-down method to explore scientific communities. They considered it useful for grant reviewers or conference planners to assess the dynamics of a scientific discipline. It could also help researchers new to a discipline, such as new PhD students or researchers looking for new publications possibilities. One of the experts also mentioned science journalists who could use CiteRiver for their journalistic inquiry. The experts stated that the prototype is fun to work with, and that they like the high degree of interactivity. One expert remarked that the approach demands some initial user training in order to grasp the data abstractions and their meaning. We agree that some initial training is necessary, but learned from our feedback sessions that the visual encodings and interactions can be learned quite fast, and experts were able to

[2] https://aclweb.org/anthology/

use the prototype after a couple of minutes of training. This also corresponds with our experience from multiple demos we gave to researchers of various backgrounds, including humanities and social sciences, who quickly learned to read and interpret the visual representations.

An interesting finding from the feedback sessions is the difference in citing behavior between the NLP and the visualization community. While the latter has a broader citation behavior and tends to cite work outside of its community, citations from the former are much more narrowly focused on NLP and data mining venues. For better orientation among the frequently cited venues, the NLP experts suggested to highlight uncommon or otherwise interesting venues or clusters to steer the users' attention to them. One of the NLP experts was exploring the topics very closely, stating that thematic communities such as machine translation, parsing, or sentiment analysis can be followed quite well and that being able to adapt cluster granularity is important to find the right abstraction level for this. The two other NLP experts also mentioned that the streamgraph for the content clustering depicts the topics within the community quite nicely. Although there are hardly any citations outside of the community, different topical foci of the aggregated venues helped to better understand and disambiguate topic streams. The expert focusing on the topic streams found that the machine translation cluster includes methodologically similar techniques such as information extraction, thus capturing the evolution of this topic. He further mentioned that the authors shown for the topics fit his perception of the group structure of the field.

Although the experts found the depicted information about the publications quite comprehensive, one aspect that was missing, as remarked by four of them, was information about co-authorship relations, author affiliations, and citation relationships between authors. For the author panel (Figure 7.1 (d)), co-authorship, or affiliation information could be used to group authors within a column. In addition, dynamic graphs that depict changing collaborations between different authors can expose a lot about the developments in a field. Citation relationships between authors, on the other hand, could help to model the thematic relations between authors and groups of authors. A great challenge for the analysis of these relationships is the limited availability of large data sets that contain a comprehensive account of citation relations. There were additional comments concerning single elements of our approach. The trendiness score attracted two experts' attention, who extensively contrasted its results to other metadata aspects of the documents. This led to insights into the significance of a publication, and thus into the structure of the field. While the experts found that for some publications the results of the score are reasonable, for some others the score differed from what they had expected. This was cause for some criticism directed at the opaqueness of the score due to its level of abstraction. Although the document trend plot (Figure 7.1 (g)) offers lists

of similar publications in past and future, the reason why these publications are similar remains unclear. We agree with the criticism and plan to add a suitable representation of common terms that two publications share as an explanation for their similarity rating.

In addition to the discussed points, the users had some feedback concerning implementation details of CiteRivers. We refer the interested reader to Heimerl et al. [2016a] for details. Overall, the feedback was positive and all experts were able to discover new findings and gain new knowledge about their community. These included facts about authors and the prominence of publications, topical developments, and their citations into the same and other communities. We can thus say, that the CiteRivers approach is effective in supporting sensemaking of scientific literature data sets and facilitates the stimulation of thoughts about the data. This helps users to derive new insights and leads to interesting discoveries that help users learn about and understand a scientific community.

7.7 Future Directions

In this chapter, we have discussed a new VA approach for the exploration of scientific literature. It comprises a novel technique to visually link grouped article sets with a user-steered abstraction of cited venues. It further integrates multiple data sources and uses them to create meaningful abstractions of the data based on several data processing methods. We demonstrated the capabilities of our approach with a use case, for which we created a corpus of visualization publications. To assess the usefulness and applicability of the approach, we asked six experts for their feedback. We find that our approach is effective for gaining insight into the thematic dynamics of a scientific field, and their relation to other communities through their citations.

There are several ways in which this line of research on visual interactive methods for scientific literature exploration can be furthered in the future. We discuss two of them here. First, in this approach, we visually linked citations and topics of a community. There is another important aspects of scientific literature that we do not include, which are author collaborations. An interesting research question in this respect is the connection between topic dynamics and their correlation to author collaboration graphs. Here, visual interactive approaches can support users in answering questions such as: How do authors collaborate across fields, and how do collaborators of single authors and groups change depending on the topic they collaborate on? Second, a similar approach to the presented one would be interesting to explore a different kind of scientific documents, namely patents. While patents have many common properties with scientific publications, they also

differ in important aspects. Two of the challenges in that respect are their specific type of language, as well as their huge quantities, demanding an effective means to interactively and visually identify interesting sets for analysis.

CHAPTER 8

Taxonomy of Approaches

This chapter contains previously published material from the following publications:

P. Federico, F. Heimerl, S. Koch, and S. Miksch. A survey on visual approaches for analyzing scientific literature and patents. *Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. doi: 10.1109/TVCG.2016.2610422

In this chapter, we introduce our classification scheme for visual interactive text analysis approaches in the scientific literature domain. We have based this scheme on a previously published taxonomy of exploration tasks [Federico et al., 2016] supported by the various approaches that exist. After describing the classification scheme and its extensions, we apply it to selected approaches from this thesis, and from the general visualization literature. The proposed extensions are based on the lessons learned from the approaches discussed in this thesis. Based on this, we sketch the current state-of-the-art in the field, and sketch paths for future research.

8.1 Classification Scheme

The nature and extent to which users are able to derive knowledge from, influence, and use NLP algorithms intuitively for interactive data exploration determines the types and quality of insights that can be generated with a given VA approach. As a categorization for this aspect of systems for scientific literature exploration and analysis, we use a previously created taxonomy for visualization approaches. It is based on abstract task definitions that approaches are designed to support. For this, we include the tasks defined by Federico et al. [2016]. The taxonomy was inspired by Shneiderman [1996]'s task by type classification. All approaches are consequently classified according to two dimensions, data types and the tasks that the exploration and analysis methods support. The data types that play a role for scientific literature analysis are *text*, *citations*, *authors*, and *metadata*. In the context of this thesis, we focus mainly on the exploration of text data from scientific literature, or its combination with other document data. For this reason, we concentrate on the text data type of the taxonomy and create a more fine-grained classification scheme for this subset of approaches. Although we focus on scientific literature in this thesis, we extend the scope of this classification to more general approaches for text analysis and exploration. We argue that most interactive visual approaches for text analysis are also applicable and useful for scientific literature data sets. Advances in the state-of-the-art of the text visualization domain in general thus also advance scientific literature text analysis in particular. This is not to say that no specialized approaches are needed for scientific literature, as has been demonstrated in this thesis. Text, however, is an important part of scientific literature, and this text analysis methods are a vital component for its exploration and analysis.

8.1.1 Task Taxonomy

In this section, we list the tasks that we used previously for the classification of visualization approaches to scientific literature. The framework is inspired by the task classification of Andrienko and Andrienko [2006], but modeled specifically to the requirements of scientific literature. It classifies interactive visualization approaches according to abstract tasks that are supported to help human users gain knowledge about certain aspects of a text data set and the documents and information it contains.

Elementary Lookup and Comparison

Approaches in this category are designed to support users in gaining insights and knowledge about single entities. Intermediate goals of such exploration sessions are,

for example, the identification of relevant entities from a larger set, or the collection and accumulation of information about a specific entity. The entities can be of many types. For scientific literature data sets these include but are not limited to topics, authors, conferences and journals. Approaches in this category not only include visualizations that depict one single entity at a time, but can also show more than one, to enable comparison between them. The joint characteristic of the approaches, however, is that they support exploration and analysis tasks that aim at single data entities. For example, visual document retrieval techniques often comprise a visualization of result sets. They are designed to compare results and identify documents relevant to the user's interest during the exploration process.

Elementary Relation Seeking

Another kind of analysis goals are those that focus on relations between entities. The approaches in this category support users with identifying entities that have a specific relation, such as publications that cite a common set of literature. In addition, such approaches can facilitate the exploration of relations according to different aspects. This includes, e.g., the exploration and identification of similarities in content between two publications whose citations overlap. Relations supported by the approaches in this category can be between entities of the same type, or of different types. Examples of the former include the analysis of author collaborations, while examples of the latter include approaches to analyze the relation between documents and the technical concepts they contain.

Patterns or Synoptic Tasks

Part of this category are approaches that focus neither on single entities, nor on local relations between some of them, but rather on the entire data set. This includes visual methods that focus on global patterns, such as finding and depicting cluster structures. These give users insights into the overall contents and arrangement of entities within a data set. This comprises the identification and comparison of such patterns. An example are collaboration circles of scientists. Here, visual interactive approaches facilitate the identification and exploration of such circles and their scientific output, including central and peripheral authors. Most prominently, science maps [Börner, 2010] fall into this category. They show scientific disciplines by grouping publications mostly according to citation links, but also based on publication content. Depending on the data set used, this results in groups of documents that represent entire disciplines of science and reveal patterns of interconnectedness.

Temporal Dynamics and Patterns

A number of approaches for visual interactive text and scientific document analysis emphasize temporal aspects of the data. They thus facilitate the exploration of and knowledge generation about its temporal dynamics. As this is one of the central aspects of many such approaches, we created a distinct category for them. We decided to do this despite the fact that this category may have some overlap with previous ones. This is due to the fact that for some approaches, although the temporal dimension of the data is at the center of exploration, some of the other tasks may be supported, too. However, the notion of time is of central importance for the analysis of scientific literature, as all progress and novel ideas are embedded into their context and the prevalent scientific ideas at that time. Examples for such approaches, that focus on temporal exploration, include those that facilitate analysis of topic dynamics and topic interaction over time. The visualization techniques in this category can be split into two subgroups. Some of them map time to space, which can be achieved by introducing a dedicated axis for the temporal aspect of entities. The other type of visualizations map time to time resulting in an animation that conveys the temporal dynamics of a data set.

8.1.2 Extension of the Taxonomy

This classification scheme can be applied to a wide range of approaches, including the ones described in this thesis. However, to be able to identify overarching properties of the approaches discussed, and to better classify visual text analysis methods in the scientific literature domain, we extend the classification of text approaches in the following according to several aspects. This not only helps to identify uncharted territory in this particular field and create new research questions. It also allows to gain useful knowledge about the collaboration of humans and machines for effective text analysis. One of the hallmarks of VA is the integration of human intentionality and cognition with automatic pattern recognition and data mining algorithms. In case of visual text exploration approaches, this primarily pertains to NLP algorithms, but may also include other techniques, such as, e.g., the extraction of cited communities described in Section 7. This integration can have different forms, and can thus give users varying degrees of interaction and controlling abilities. Inspired by the classification described by Bertini and Lalanne [2009], we have created categories for scientific literature approaches that are also applicable to a broader range of interactive text visualization methods. This is demonstrated in Section 8.2. The categories help to order and classify existing algorithms, identify interesting patterns that give insight into some of the constraints that govern approaches currently published, and provides new ideas or future research endeavors. We identified four modes of human control

over algorithms currently employed in existing interactive visual text analysis approaches. Those are listed and discussed in the following.

Static

This category describes algorithm integration that processes data entirely autonomous from any user intervention. Usually, the designer or developer of the system anticipated some data mining algorithm that summarizes or abstracts data for users. An example for this are keyword extraction algorithms that are implemented in a static way such that the algorithm runs automatically, and keywords are extracted and presented to the user for interpretation. Another example is the application of a clustering algorithm to document data that operates with a fixed set of parameters and yields a fixed number of clusters that are presented to users. An example for this is, e.g., the publication trendiness for CiteRivers that is computed for every document in the data set and provides users with the results for interpretation (see Chapter 7).

Granularity

These types of integration approaches are in many ways similar to the static scheme in that users cannot influence the processing of the data in any way. Exploring the results of this processing, however, can be adapted to different exploration styles and foci depending on the information that users are interested in. This is achieved by allowing to interactively adapt the granularity of the results provided by NLP processing algorithms. An example for this is letting users define the number of clusters for a given clustering algorithm interactively. This is, for example, supported for the topic stream and the cited communities in the CiteRivers approach from Chapter 7. In addition, hierarchical topic modeling approaches for visual text exploration often use this type of integration, e.g., Dou et al. [2013], who allow users to split, join, and merge extracted topics.

Scope

With this mode of integration, users are able to influence NLP algorithms by defining the set of instances that they are applied on. There are a plethora of possible interaction methods with which such an integration could be achieved. For example, users could be asked to defined or manipulate visually represented queries that, e.g., consist of keywords of documents. The resulting set of texts is then processed by an algorithm. This is, for example, what Koch et al. [2011a] support with their system. Another instance of this integration paradigm are the text characterization algorithms of DocuCompass from Chapter 4. In this approach, users define the exploration granularity and focus on a document set. Based on

this interaction, one or multiple algorithms for text characterization are applied to the focused set, and the results are displayed to the user.

Steering

Steering is the closest coupling of human users and pattern discovering algorithms. In this mode of integration, users are enabled, through suitable interaction methods, to influence statistical models of the data used for the analysis. This can either be done by directly allowing users to modify and influence hypotheses parameters, or by labeling or manipulating data instances to provide feed back to the algorithm. Endert et al. [2011] call these two types of interactions *parameter level* and *observation level* interaction, respectively. An example for steering is the classifier creation approach that we present in Chapter 5. Other examples from the visualization literature is the approach presented by Endert et al. [2012].

There is another aspect according to which the integration of data mining and NLP methods into interactive visualization approaches has been classified previously. Bertini and Lalanne [2009] dub their concepts *white box* and *black box* integration. They describe white box integration as a scheme that provides visualizations of the inner workings of the algorithms and corresponding models. In contrast, black box integration hides this information from users, and only provides insight into the models through its effect on the data instances depicted. We consider this concept orthogonal to our taxonomy, which describes the influence users have on algorithms through interaction. Perfect white box integration is rare, a fact that is also acknowledge by Bertini and Lalanne [2009]. One example is van den Elzen and van Wijk [2011], who provide an interactive visualization that gives users full insight and control over a decision tree construction algorithm. In addition, our approach from Chapters 5 and 6 are examples of partial white box approaches. An example for black box integration is, e.g., the keyword extraction algorithm from the CiteRivers approach described in Chapter 7. Mühlbacher et al. [2014] discuss the integration of data mining algorithms into visualization systems. They take a technical stance and discuss the integration of implemented algorithms into visualization systems to support various types of user integration.

So far, we have created a taxonomy with two dimensions. It organizes existing approaches of interactive visual text analysis according to the abstract tasks supported, and the integration of the machine learning and NLP methods that help users produce knowledge about data sets and their contents. However, there are further categories that are useful when grouping and organizing approaches. Firstly, the categories so far define abstract tasks of human users that are supported by interactive visualizations and the way that users can interact with these visualizations. However, we have not yet specified the user group that is targeted by

an approach. According to the two-level model presented in Chapter 2, we have two user types that correlate with the two levels of the model. Those are the *domain* expert, that wants to explore data sets to gain knowledge about them, and the *linguist*, who wants to analyze language data to improve NLP processing methods.

After defining categories for human analysis tasks and the integration of humans and automatic algorithms, one important aspect is still missing from the taxonomy. It is the classification of the language processing algorithms that can be included into a system to help humans with sensemaking, exploration, and data analysis. Categorizing the methods helps to determine which types of information extraction tasks are possible with a given approach, and how suitable visualization and interaction methods have to be designed to include human users into these processes. For this, we have devised a classification that is based on the different knowledge fields of linguistics [Jurafsky and Martin, 2009] that are relevant for written texts. For all of these fields, suitable NLP methods exist and are accessible, e.g., through libraries such as CoreNLP [Manning et al., 2014]. We list and discuss these categories in the following.

Token

This is not a field of linguistics, but this processing depth is popular in visualization approaches, and we thus include it as a separate category. It is based on the extraction of single tokens from texts according to the bag-of-words model described in Chapter 2. The extraction is typically based on rule systems that extract single tokens separated by white space, and identifies sentence boundaries based on punctuation. In addition, stop word lists are frequently used to remove unwanted words. Then, statistic methods are usually applied to the resulting tokens. This yields word counts that can then, e.g., be depicted as word clouds, which we do in Chapter 3. Alternatively, more complex statistical schemes, such as the popular LDA algorithm [Blei et al., 2003], that clusters tokens into topics, can be applied. In addition, methods such as NER are applied at the token level. They decide whether a token is part of a named entity based on its local context. Interaction with the processing component can have many different forms on the token level, from interactively adding and removing terms for the stop word list to interactively steering complex topic modeling algorithms.

Morphology

Morphology describes the composition of words from smaller, meaningful entities [Jurafsky and Martin, 2009] which include, e.g., affixes in English. Morphological processes also govern the inflection of nouns or conjugations of verbs.

Components for morphological processing in visualization approaches are typically either stemmers or lemmatizers. While the former are made up of a set of rules that are applied to tokens to remove suffixes, and thus yield the stem of a word, the latter tries to find the correct base form. Due to the rules being applied successively on the token strings, stemming is a very fast process. Its results, however, can be hard to interpret by humans, as they are often non-existing words from which suffixes have been chopped off. Lemmatization, in contrast, reduces tokens to their root forms or lemmas, such as *saw* → *see*. It typically includes deeper morphological knowledge and is thus capable of handling irregular word forms. Burch et al. [2013] provide a visual text exploration approach that joins tokens in word clouds according to their stems, using a scheme similar to stemming. In addition, stemming or lemmatization can be used to improve the extraction of keywords from texts, as we do in Chapters 3 and 4.

Syntax

Syntactic theories describe the combination of single words to larger structures, such as sentences. A natural language parser is able to uncover these hidden structures. It thus provides information about the relation of words or phrases within these larger structures. This knowledge is important for many NLP tasks, such as machine translation, opinion analysis, or language generation. There are two families of theories in the field of syntax, those that are based on sentence constituents, and those that assume dependency relations between words. Both have lead to different formalisms to describe syntactic structures and processes. While the former stipulate the existence of constituents that subsume single words into larger units, the latter is based on binary relations between single words. Syntactic structures of the former type are thus well-formed trees, while dependency structures lead to directed acyclic graphs to describe word relations within a sentence. In visualization approaches for natural language texts, syntactic information has been used in various contexts. Oelke et al. [2010] use it as a source of information for the visualization of sentence complexities in longer documents. Our approach for feature exploration from Chapter 6 uses it for the classification of citations in scientific publications. Collins et al. [2009b] visualize syntactic information to allow NLP experts to detect and solve problems with their machine translation models.

Semantics

Semantics studies the meaning of single words in a text, and how these meanings combine within sentences. There are many different theories for semantic analysis. Semantics includes the study of phenomena such as homonymy and polysemy [God-

dard, 2011] that can be relevant when, e.g., extracting single words in token-based processing to depict to users. Automatic implementations of semantic theories into processing resources are typically experimental and exhibit low accuracy. For this reason, techniques such as semantic role labeling or frame semantic parsers are rarely used as part of VA systems. There are, however, dictionary projects that capture semantic relations between words such as *pants* and *jeans*, combining them to synonym groups. This has, for example been used in DocuBurst [Collins et al., 2009a] that depicts hyponymy relations between terms extracted from a document. Attempts have also been made to create 3D scenes based on their semantic representation of their natural language description [Coyne and Sproat, 2001]. Recently, Kurzhals et al. [2016] have presented an approach that uses semantic information from movie scripts to compute movie scene similarities.

There are two further knowledge fields of linguistics, namely pragmatics and discourse linguistics. While the former deals with the meaning of natural language expressions that depend on context and the speaker's intention [Jurafsky and Martin, 2009; Levinson, 1983], the latter focuses on language structures beyond the scope of single sentences [Jurafsky and Martin, 2009]. However, these fields have not played any role in current text VA approaches, probably due to the lack of high-accuracy scalable implementations. The only exception, to the best of our knowledge, are Zhao et al. [2012], who visualize the tree structure output of a discourse parser to help linguists analyze them.

Recently, a new type of methods has become popular in NLP that operates mostly on the token level. Based on this, they are able to learn and model complex linguistic phenomena from large training corpora. These methods, that go by the name of *deep learning*, employ multi-level representations of knowledge that allow them to encode complex dependencies and relations in their models. While for these techniques, the depth of linguistic processing becomes less important, as they function on raw token input, user integration becomes a problem of even greater importance. Not only to allow users to learn about the inner workings of the algorithms, but also to apply these techniques as part of interactive data exploration and analysis approaches. For such deep learning techniques, first visual interactive methods started appearing recently [Chuang and Socher, 2014; Smilkov et al., 2016; Liu et al., 2016]. These focus primarily on debugging problems and understanding the information encoded within complex models. This is an important task for linguists that develop and improve such techniques. On the other hand, for domain experts, who want to explore and analyze a text data set, these well performing algorithms can help with a wide range of analysis and exploration tasks. For both scenarios, good visual representations and interactive exploration and analysis approaches are a challenge for future endeavors in text VA.

8.2 Application of the Extended Scheme

In the following, we apply the previously discussed taxonomy to approaches from the visualization literature. We include all of our own approaches, and classify them according to our scheme. In addition, we add multiple other approaches to the categories to show the general applicability of the taxonomy. This also provides insight into what areas in the entire space are well covered by current research, and where new research opportunities lie. If there are more approaches that fit or list, we select the latest ones. In addition, we also mention multiple older ones if we consider them relevant for the subsequent discussion. This means that we are not aiming to create a comprehensive overview of the current state-of-the-art of interactive text visualization approaches. Instead, we aim to demonstrate the applicability of the taxonomy that is rooted within the projects pursued as part of this thesis. For a survey of text-based and other visualization approaches to scientific literature we refer readers to Federico et al. [2016].

The 20 example approaches we selected to exemplify the taxonomy are all listed in Table 8.1. We have decided to use the integration of users and NLP processing methods as the primary dimension for organizing the approaches, as this is the dimension along which the approaches discussed within this thesis exhibit the most variance. As a consequence, we have aligned this dimension vertically in Table 8.1, grouping the approaches according to it. Some approaches include multiple NLP methods whose integration may pertain to different categories. This is, e.g., the case for our approach from Chapter 7, which includes static keyword extraction methods, and a clustering technique whose results can be explored on different levels of granularity. In such a case, we list the approach in the category with the highest level of user control that users have for each of the available methods. The other, secondary dimensions are included as columns of the table, each cell indicating whether a certain attribute holds for the approach that occupies the respective line of the table. Those dimensions are the *level* of the two-level model (see Chapter 2), the abstract analysis *tasks* discussed above, and the processing level of the NLP method based on the respective knowledge field of *linguistics*. Despite the fact that they are exclusive for many approaches, more than one of some of the secondary dimensions can hold for a couple of examples. This is true, e.g., for the *temporal* task, as it is often combined with other tasks in existing VA systems. A fact that has also been discussed in Federico et al. [2016].

In the following, we will go through the primary dimensions and mention all of the examples, briefly discussing the reasons for their classification. The *static* category contains approaches that apply NLP methods statically to text data, and no user control at all is possible. These approaches mostly use text data as a resource for keywords from which static labels are created in the visualizations. The first

	approach	domain	linguist	elementary	relative	synoptic	temporal	token	morphology	syntax	semantics
		level		tasks				linguistics			
static	Chen [2006]	●	○	○	○	●	○	●	○	○	○
	Collins et al. [2009b]	○	●	○	●	○	○	○	○	●	○
	van Ham et al. [2009]	●	○	○	○	●	○	●	○	●	○
	Burch et al. [2013]	●	○	●	○	○	○	○	●	○	○
	Fried and Kobourov [2014]	●	○	○	○	●	○	●	○	○	○
granularity	Collins et al. [2009a]	●	○	○	●	●	○	○	○	○	●
	Oelke et al. [2010]	●	○	○	○	●	○	○	○	●	○
	Dou et al. [2013]	●	○	○	○	●	●	●	○	○	○
	Chapter 3	●	○	○	○	●	○	○	●	○	○
	Chapter 7	●	○	○	○	●	●	○	●	○	○
scope	MacEachren et al. [2011]	●	○	○	○	●	●	●	○	○	○
	Bosch et al. [2013]	●	○	○	○	●	●	○	●	○	○
	Chapter 4	●	○	○	○	●	○	○	●	○	○
	Kim et al. [2016]	●	○	○	○	●	○	○	●	○	○
	Zhang et al. [2016]	●	○	○	●	○	●	○	○	○	●
steering	Chapter 5	●	○	○	●	○	○	○	●	○	○
	Chapter 6	○	●	○	○	●	○	○	○	●	○
	Endert et al. [2012]	●	○	○	○	●	○	●	○	○	○
	Choo et al. [2013]	●	○	○	○	●	○	●	○	○	○
	Bradel et al. [2014]	●	○	○	○	●	○	●	○	○	○

Table 8.1 — Matrix of our taxonomy for interactive visual text analysis approaches for scientific literature and other domains. For each mode of the modes of human control, we list five example approaches from this thesis and the general visualization literature, with a strong bias towards more recent examples.

approach in this category is Cite Space II [Chen, 2006]. It visualizes co-author graphs that allow for the analysis of research dynamics in a field. As it provides a comprehensive picture of a scientific field, it is part of the *synoptic* category. Keywords from the publication abstracts are used to provide meaningful labels for the graph nodes. Collins et al. [2009b] apply their BubbleSets technique to multiple data types, including syntactic parse trees. Here, BubbleSets are used to depict relations between syntactic constituents of English and Chinese sentences in a machine translation model. This gives linguists insights into their data and models that allows them to understand and solve problems. van Ham et al. [2009] visualize word sequences in a graph, showing different co-occurrences of words within a corpus. They present two version of their approach, one based on syntactic structures, and the other one based on token sequences and regular expressions. Users cannot view the entire data set at once, but have to select patterns or create new ones. These are shown at a single granularity level. Burch et al. [2013] show texts as word clouds that combine words with identical prefixes, uncovering their morphological structure to some extent. Fried and Kobourov [2014] extract keywords from the publication titles in a large scientific literature data set and organize them into a graph. This graph is laid out in 2D space, serving as the basis for a science map that is clustered into different thematic regions.

The next category of approaches, *granularity*, allows users to browse and explore static algorithmic results on different levels of granularity. Collins et al. [2009a] introduce a visualization of hyponymy relations of a document set that users can explore with different foci and at different levels of granularity. It can also be used to compare different text bodies with respect to the concepts that are mentioned within them. Oelke et al. [2010] allows users to explore the readability of text bodies on different levels of granularity, including entire documents and single sentences. Their readability measure is based on different types of linguistic information, including syntactic structure of the sentences. Dou et al. [2013] create a topic hierarchy of a text data set whose temporal dynamics can be explored and modified by interacting with a tree visualization. Our approaches from Chapters 3 and 7 both allow users to view and browse word clouds and topic streams, respectively, at various levels of granularity.

The approaches in the *scope* category provide users with interactive control of the scope of the data that NLP methods are applied to. MacEachren et al. [2011] and Bosch et al. [2013] introduce approaches for situational awareness based on interactive visual micro blog message analysis. The former allow filtering according to multiple aspects, including location and time, and extract named entities from the retrieved messages that can be searched by users. The latter provide a flexible, graph-based, visual filter orchestration method that allows to interactively combine various types of message filters. Both systems facilitate synoptic tasks

and emphasize the temporal dimension of the data. The lens-based approach from Chapter 4 allows users to apply NLP methods to arbitrary regions of a document spatializations during exploration. Kim et al. [2016] provide a comparable but less flexible approach based on one particular topic modeling algorithm. Zhang et al. [2016] uses semantic role labeling to match message contents to a list of predefined message categories. This is achieved by comparing the predicate argument structure of the messages to example phrases for the categories. Users can select the scope of messages based on a geographic lens, allowing them to explore the messages of different categories in the selected area.

Users can give feedback or influence the parameters of NLP resources through interaction in the visual methods of the *steering* category. Chapters 5 and 6 introduce approaches that let users explore and improve classification methods for text data. The former is designed for domain experts and supports exploring and learning about the differences between two text categories. The latter can be used by linguists to create classification approaches for language data. It supports linguistic data of various depth, up to the syntactic level in the example of Chapter 6. Endert et al. [2012] introduces the concept of semantic interaction that allows users to give feedback by changing the distance between document glyphs in a spatialization. This changes and updates the similarity computation of the underlying text spatialization model. Document similarity in this approach is defined as the named entity overlap of two documents. User feedback does not influence the underlying NLP facilities, but rather the computation of document similarities based on NER results. Bradel et al. [2014] extend the approach by gauging the user's interest based on her feedback, and retrieve additional, potentially relevant, documents and add them to the spatialization. Choo et al. [2013] discuss an approach that lets users interactively shape a topic model according to their wishes through interacting with a document spatialization and its keyword labels.

8.2.1 Discussion

Although Table 8.1 is far from being comprehensive, it gives a good overview of the research directions that have been pursued in interactive visual text analysis over the past years. For this reason, we base a succinct discussion of the current state-of-the-art and potential future work on it. The first thing that becomes apparent when looking at Table 8.1 is that support for linguistic data analysis for NLP experts (see also the two level model in Chapter 2) is particularly scarce. Here, future research opportunities lie in two different areas that both pave the way for better and more targeted NLP methods for interactive visual text exploration and analysis approaches. Firstly, language data in general can be made explorable and methods can be devised that help linguists analyze and learn from their data.

This would be especially beneficial in data driven linguistic disciplines, such as corpus linguistics. Recently, first research into this direction has appeared, e.g., by Schneider et al. [2016] and Köper et al. [2016], and is driven by visualization researchers and linguists alike under the term *visual linguistics*. However, there are still many open challenges for the use of interactive visualization methods to further linguistic research. This includes visual representation for complex linguistic data and statistical patterns of, e.g., usage scenarios, and corresponding interaction methods. One example for this would be tree banks that contain large numbers of syntactic trees. Here, customized visualization schemes could help to provide linguists with an overview of prominent syntactic phenomena, and drill down on relevant information with suitable interaction. Secondly, specialized interactive approaches can help linguists to create customized NLP methods for specific exploration and analysis scenarios. An example for this is the approach from Chapter 6 that helps create classification approaches for language data. Many other scenarios are thinkable that could be supported, including, e.g., the identification and classification of specific kind of named entities, or specific semantic relations between mentioned persons in texts.

However, not only adapting or customizing existing methods is a relevant research direction for the field of text VA. Currently, existing NLP methods are mostly used within text VA approaches, often with off-the-shelve implementations. Here, new specialized NLP methods designed for interactive analysis of language data are a future research challenge for visualization and NLP researchers alike. The reason for this is that processing resources used within interactive systems have largely differing requirements, compared to those used to autonomously process text. This includes scalability and processing speed, as illustrated by van Ham et al. [2009], who deem natural language parsing unsuitable for interactive analysis in their approach. Towards these ends, the linguistic competence of human users can be leveraged by providing efficient processing resources that offer support to users with interpreting language data, rather than entirely deciphering linguistic structures on their own. In addition, with increasing degree of user inclusion, interpretability of the results, as well as intermediate processing steps are of importance. Here, new visual encodings and metaphors are needed, as well as suitable interactions methods that translate interaction with these visualizations into feedback for algorithms. This is akin to what Endert et al. [2012] propose for document spatializations at the token level of linguistic structure. However, users have to be aware of the changes and adaptive processes of the models based on their feedback. Linguists and visualization researchers will have to closely collaborate to create effective approaches, with the former focusing on scalability and interpretability, and the latter providing adequate visual representations and suitable interaction methods. Effective NLP methods do not always have to involve

complex data mining algorithms that are equipped with linguistic information, but can also be based on hand curated resources or dictionaries that help to organize, categorize, or abstract language data, as Collins et al. [2009a] and Zhang et al. [2016] effectively demonstrate.

Another interesting fact from Table 8.1 is that the majority of approaches in the *scope* category deal with the interactive visual analysis of micro blog data in their geographic context. This is not surprising, given the fact that situational awareness, which these approaches aim to facilitate, is maintained through geographically dependent analysis and exploration tasks. Here, the application of NLP methods that extract useful information and patterns from collections of messages can be adapted to the geographical location, the temporal dimensions, and other aspects that the user knows or suspects to contain relevant information. Interestingly, the remaining two approaches are methods that translate this geographical metaphor to textual documents that are laid out according to document similarities or other data aspects. Here, users are supported with a highly flexible interaction technique that, through its many degrees of freedom it offers to users, exhibits a certain playfulness that is conducive to human sensemaking (see Chapter 2). In particular, such approaches are useful for scientific literature analysis, because it is a paramount tool to combine and correlate different types of data through multiple types of visualizations suitable to them.

For scientific literature analysis, correlation and integration of various types of data is important to enable users to explore and analyze publications within their scientific context (see also Chapter 7). Being able to apply text processing methods at various degree of granularity and in a playful and highly flexible manner can help to analyze, e.g., co-author graphs, and scientific collaboration networks, visualization of citation relations, and many more. Suitable focus+context techniques, such as the discussed magic lens based ones, can help users understand these visualization of publication meta data in the context of their contents by enabling the application of text processing methods to confined regions of a visualization and get context-dependent views on the text data. Future research along these lines can thus help to advance interactive visualizations for scientific literature analysis and exploration.

Chapter 9

Conclusion

In this thesis, we have presented and discussed several approaches that have contributed to interactive visual text analysis. To be able to accurately describe these approaches, we have introduced the two level model of text VA in addition to discussing relevant foundations and background from the field of VA as well as NLP. Despite the fact that many of our approaches can be applied to general text data sets, we have a particular focus on the analysis of scientific literature which was illustrated by respective use cases and example data sets.

Word Cloud Explorer, the first approach discussed in this thesis is designed for the interactive exploration of single, coherent text bodies. It is based on word clouds as the main visualization paradigm and offers advanced interaction possibilities that facilitate an effective drill-down to information relevant during exploration. This makes the approach suitable for casual users that can start to explore data without much initial training. We further extended the approach by designing a visual representation for multiple documents based on similar principles. Contrastive exploration of documents based on the word cloud paradigm thus becomes possible. We have demonstrated this for literary texts as well as technical documents. For the free exploration of larger document data sets, we have presented DocuCompass, a flexible interaction technique for 2D document spatializations. It can be combined with arbitrary types of spatial layouts and applied to many types of documents from different domains. The technique is based on magic lenses that allow to flexibly apply text and meta data characterization methods on documents in arbitrarily sized subsets of a spatialized document set. Depending on the configuration, it

exhibits a particularly high scalability with respect to the size of the document set, and heterogeneity of content.

When users get interested in particular aspects of a data set during exploration, search and retrieval tasks become relevant to filter for interesting information. To this end, we discussed an approach that allows users to store knowledge about a document data set gained during exploration. It supports the exploration of the state of a text classifier in the context of a text data set. During this process, the design of the interactive approach fosters mutual learning between the user and the classification algorithm. The resulting classifier can be stored to filter future data, e.g., in dynamic data sets. Alternatively, results can be re-evaluated and updated any time within the interactive environment. In addition to domain experts, who can use ML approaches for classification and filtering of text data, linguists and NLP specialists work on machine learning-based classification approaches to process language data and create automated analyses. Here, creating features to accurately describe language data is an important task. The next approach, FeatureForge, supports linguists with this task by providing them with interactive visual tools to make sense of their data. It focuses particularly on labels and feature representations of linguistic entities. The resulting classifiers and language processing methods can then be used for text mining as part of VA approaches.

The final approach of this thesis tackles an analysis problem that is specific to scientific literature and related text types, such as patents. These types of documents are embedded into a scientific context that needs to be analyzed in conjunction with document contents to accurately capture the entire data set. With the CiteRivers approach, we support exploring contents and trendy topics of scientific literature data sets. This information is visually linked with the scientific surroundings of these publications based on their citations. We specifically focus on document meta data, including citation links, to provide information about a document's context. The resulting visual interactive approach provides a more holistic view of a document data set for interactive exploration than previous approaches.

Finally, based on the lessons learned during the design of our approaches, we extended a taxonomy for scientific literature analysis and exploration approaches to general text visualization. We introduced several aspects according to which approaches can be classified. Moreover, the taxonomy's applicability beyond the scope of the thesis is demonstrated by applying it to a selection of recent publications from the text visualization literature. Based on these representative examples of the current state-of-the-art, we derive ideas and visions for future research endeavors in the field. These comprise new methods that help linguists explore and analyze their language data and the development of specialized techniques to support human sensemaking rather than the autonomous processing of natural

language. In addition, flexible interaction techniques are necessary that allow users to aggregate and analyze text data efficiently at various levels of granularity and in a context-dependent fashion. This facilitates the integrated analysis of meta data rich document sets, such as scientific literature sets. To foster research on such approaches, we have further compiled and updated a data set of visualization publications while working on the projects of this thesis, which we published as part of a larger data set.

This thesis contributes to VA, a field that is concerned with the interactive visual analysis of large and complex data collections. As such, it becomes more and more relevant in today's increasingly computerized world. On the one hand, computerization leads to ever larger sets of digitally stored data containing information that can be leveraged through the combination of automatic processing and interactive visualization. This has been demonstrated for text data sets mostly from the scientific literature domain throughout this thesis. However, even in cases in which information from data is extracted and leveraged by fully automatic means, VA helps humans understand and, if possible, control such processes. From autonomous vehicles and machine translation to advanced manufacturing, VA is a vital approach that helps to satisfy one of the most basic human traits in an increasingly complex and automated world: the desire to learn and understand.

Bibliography

S. Abney. Statistical methods and linguistics. *The balancing act: Combining symbolic and statistical approaches to language*, pages 1–26, 1996. [page 24]

S. Abney. *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC, 1st edition, 2007. ISBN 978-1-58488-559-7. [pages 29, 31, 32, and 33]

S. F. Adafre and M. de Rijke. Feature engineering and post-processing for temporal expression recognition using conditional random fields. In *ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing (FeatureEng)*, pages 9–16, 2005. [page 125]

A. Ahmed, T. Dwyer, C. Murray, L. Song, and Y. X. Wu. InfoVis 2004 Contest: Wilmascope graph visualisation. In *Symposium on Information Visualization (InfoVis)*. IEEE, 2004. [pages 43 and 140]

E. Alexander and M. Gleicher. Task-driven comparison of topic models. *Transactions on Visualization and Computer Graphics*, 22(1):320–329, Jan 2016. doi: 10.1109/TVCG.2015.2467618. [page 39]

E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher. Serendip: Topic model-driven visual exploration of text corpora. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182. IEEE, Oct 2014. doi: 10.1109/VAST.2014.7042493. [pages 42, 66, and 70]

E. Alexander, C.-C. Chang, M. Shimabukuro, S. Franconeri, C. Collins, and M. Gleicher. The biasing effect of word length in font size encodings. In *Visualization Posters*. IEEE, Oct 2016. [pages 48 and 52]

B. Alsallakh, A. Hanbury, H. Hauser, S. Miksch, and A. Rauber. Visual methods for analyzing probabilistic classification data. *Transactions on Visualization and Computer Graphics*, 20(12):1703–1712, 2014. [page 99]

N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer, Berlin, 2006. ISBN 978-3-540-25994-7. [page 162]

M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: An interactive approach to decision tree construction. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 392–396. ACM, 1999. [page 99]

M. Ankerst, M. Ester, and H.-P. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 179–188. ACM, 2000. doi: 10.1145/347090.347124. [page 98]

J. L. Austin. *How to do things with words*. William James Lectures. Harvard University Press, 2nd edition, 1975. ISBN 0-674-41152-8. [page 1]

S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Conference on Hypertext and Hypermedia (HT)*, pages 193–202. ACM, 2008. doi: 10.1145/1379092.1379130. [page 48]

M. J. Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989. [page 17]

F. Beck, S. Koch, and D. Weiskopf. Visual analysis and dissemination of scientific literature collections with SurVis. *Transactions on Visualization and Computer Graphics*, 22(01), January 2016. [page 140]

M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 43–52. IEEE, 2014. [page 99]

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. [page 135]

G. Berend and R. Farkas. Sztergak: Feature engineering for keyphrase extraction. In *International Workshop on Semantic Evaluation (SemEval)*, pages 186–189, 2010. [page 125]

E. Bertini and D. Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *SIGKDD Workshop on Visual Analytics and Knowledge Discovery (VAKD)*, pages 12–20. ACM, 2009. doi: 10.1145/1562849.1562851. [pages 98, 164, and 166]

E. Bertini, M. Rigamonti, and D. Lalanne. Extended excentric labeling. *Computer Graphics Forum*, 28(3):927–934, 2009. doi: 10.1111/j.1467-8659.2009.01456.x. [pages 71 and 74]

E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and magic lenses: The see-through interface. In *Annual Conference on Computer*

Graphics and Interactive Techniques (SIGGRAPH), pages 73–80, New York, NY, USA, 1993. ACM. doi: 10.1145/166117.166126. [page 71]

S. Bird, R. Dale, B. Dorr, B. Gibson, M. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. Radev, and Y. F. Tan. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1755–1759. European Language Resources Association, 2008. [page 131]

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, Mar. 2003. [pages 38 and 167]

K. Börner. *Atlas of Science: Visualizing What We Know*. The MIT Press, 2010. ISBN 0262014459, 9780262014458. [pages 39 and 163]

H. Bosch. *Casual analytics: advancing interactive visualization by domain knowledge*. PhD thesis, University of Stuttgart, 2014. [page 5]

H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl. ScatterBlogs: Geo-spatial document analysis. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 309–310. IEEE, Oct 2011. doi: 10.1109/VAST.2011.6102488. [pages 68 and 71]

H. Bosch, D. Thom, F. Heimerl, E. Püttmann, S. Koch, R. Krüger, M. Wörner, and T. Ertl. ScatterBlogs2: Real-time monitoring of microblog messages through user-guided filtering. *Transactions on Visualization and Computer Graphics*, 19 (12):2022–2031, 2013. doi: 10.1109/TVCG.2013.186. [pages 7, 68, 71, 119, 121, 171, and 172]

L. Bradel, C. North, L. House, and S. Leman. Multi-model semantic interaction for text analytics. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 163–172. IEEE, Oct 2014. doi: 10.1109/VAST.2014.7042492. [pages 171 and 173]

M. Brooks, S. Amershi, B. Lee, S. M. Drucker, A. Kapoor, and P. Simard. Featureinsight: Visual support for error-driven feature ideation in text classification. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 105–112. IEEE, 2015. [page 126]

T. Bühler and M. Hein. Spectral clustering based on the graph p-laplacian. In *International Conference on Machine Learning (ICML)*, pages 81–88, New York, NY, USA, 2009. ACM. doi: 10.1145/1553374.1553385. [pages 35 and 36]

M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf. Prefix tag clouds. In *International Conference on Information Visualisation (IV)*, pages 45–50, Washington, DC, USA, 2013. IEEE. doi: 10.1109/IV.2013.5. [pages 46, 47, 49, 63, 168, 171, and 172]

M. Burch, S. Lohmann, F. Beck, N. Rodriguez, L. D. Silvestro, and D. Weiskopf. Radcloud: Visualizing multiple texts with merged word clouds. In *International Conference on Information Visualisation (IV)*, pages 108–113. IEEE, July 2014. doi: 10.1109/IV.2014.72. [page 49]

C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998. [page 31]

A. W. Burks. Peirce's theory of abduction. *Philosophy of Science*, 13(4):301–306, 1946. [page 23]

L. Byron and M. Wattenberg. Stacked graphs; geometry and aesthetics. *Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, Nov 2008. doi: 10.1109/TVCG.2008.166. [pages 140 and 141]

J. Cai and M. Strube. Evaluation metrics for end-to-end coreference resolution systems. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 28–36. Association for Computational Linguistics (ACL), 2010. [page 30]

C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *International Conference on Machine Learning (ICML)*, pages 111–118. ACM, 2000. [page 34]

S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, CA, USA, 1999. ISBN 1-55860-533-9. [pages vi, 2, 13, 14, 15, and 71]

Q. Castella and C. Sutton. Word storms: Multiples of word clouds for visual comparison of documents. In *International Conference on the World Wide Web (WWW)*, pages 665–676. ACM, 2014. [page 49]

M. Chalmers and P. Chitson. Bead: Explorations in information visualization. In *Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 330–337. ACM, 1992. [page 38]

J. Cheeger. *A lower bound for the smallest eigenvalue of the Laplacian*, pages 195–199. Princeton University Press, June 1970. ISBN 0691080763. [page 35]

C. Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35(3):401–420, 1999. [page 43]

C. Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. [pages 43, 140, 149, 171, and 172]

C. Chen, J. Zhang, W. Zhu, and M. Vogeley. Delineating the citation impact of scientific discoveries. In *Joint Conference on Digital Libraries (JCDL)*, pages 19–28. ACM/IEEE, 2007. [page 148]

Y.-X. Chen, R. Santamaría, A. Butz, and R. Therón. TagClusters: Semantic aggregation of collaborative tags beyond tagclouds. In *International Symposium on Smart Graphics (SG)*, pages 56–67. Springer, 2009. doi: 10.1007/978-3-642-02115-2_5. [page 48]

E. H.-h. Chi and J. T. Riedl. An operator interaction framework for visualization systems. In *Symposium on Information Visualization (InfoVis)*, pages 63–70. IEEE, 1998. [page 14]

N. Chomsky. *Language and Thought*. Anshen transdisciplinary lectureships in art, science, and the philosophy of culture. Moyer Bell, 1993. ISBN 9781559210744. [page 1]

J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, 2013. doi: 10.1109/TVCG.2013.212. [pages 24, 42, 70, 171, and 173]

J. Chuang and R. Socher. Interactive visualizations for deep learning. In *VIS Workshop on Visualization for Predictive Analytics*. IEEE, 2014. [page 169]

J. Chuang, C. D. Manning, and J. Heer. "Without the Clutter of Unimportant Words": Descriptive keyphrases for text visualization. *Transactions on Computer-Human Interaction*, 19(3):19:1–19:29, Oct. 2012a. [pages 29 and 36]

J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *International Working Conference on Advanced Visual Interfaces (AVI)*, pages 74–77. ACM, 2012b. [pages 41 and 66]

J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Conference on Human Factors in Computing Systems (CHI)*, pages 443–452, New York, NY, USA, 2012c. ACM. doi: 10.1145/2207676.2207738. [pages 40 and 72]

J. Chuang, D. Ramage, D. A. McFarland, C. D. Manning, and J. Heer. Large-scale examination of academic publications using statistical models. In *AVI Workshop on Supporting Asynchronous Collaboration in Visual Analytics Systems*, 2012d. [page 41]

J. Clark, C. Yallop, and J. Fletcher. *An Introduction to Phonetics and Phonology*. Blackwell Textbooks in Linguistics. Blackwell Publishing, 3rd edition, 2007. ISBN 9781405130837. [page 1]

A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *Computing Surveys*, 41(1):2:1–2:31, Jan. 2009. doi: 10.1145/1456650.1456652. [pages 68 and 71]

C. Collins. *Interactive Visualizations of Natural Language*. PhD thesis, University of Toronto, 2010. [page 14]

C. Collins, S. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009a. doi: 10.1111/j.1467-8659.2009.01439.x. [pages 71, 169, 171, 172, and 175]

C. Collins, G. Penn, and S. Carpendale. Bubble Sets: Revealing set relations with isocontours over existing visualizations. *Transactions on Visualization and Computer Graphics*, 15(6):1009–1016, 2009b. [pages 168, 171, and 172]

C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze facted text corpora. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 91–98. IEEE, 2009c. doi: 10.1109/VAST.2009.5333443. [pages 36, 38, 49, 71, and 145]

M. Correll, M. Witmore, and M. Gleicher. Exploring collections of tagged text for literary scholarship. *Computer Graphics Forum*, 30(3):731–740, 2011. doi: 10.1111/j.1467-8659.2011.01922.x. [page 70]

G. Costagliola and V. Fuccella. CyBiS: A novel interface for searching scientific documents. In *International Conference on Information Visualisation (IV)*, pages 276–281. IEEE, 2011. doi: 10.1109/IV.2011.95. [page 40]

I. G. Councill, C. L. Giles, and M.-Y. Kan. Parscit: An open-source crf reference string parsing package. In *International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2008. [page 152]

B. Coyne and R. Sproat. Wordseye: an automatic text-to-scene conversion system. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 487–496. ACM, 2001. [page 169]

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2010. ISBN 978-0-521-78019-3. [pages 32 and 33]

D. Crystal. *The Cambridge Encyclopedia of Language*. Cambridge University Press, 2nd edition, 2003. ISBN 0521559676, 0521550505. [page 1]

W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu. Context-preserving, dynamic word cloud visualization. *Computer Graphics and Applications*, 30(6): 42–53, 2010. doi: 10.1109/MCG.2010.102. [page 49]

W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, 2011. [page 140]

W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *Transactions on Visualization and Computer Graphics*, 20(12): 2281–2290, Dec 2014. doi: 10.1109/TVCG.2014.2346433. [pages 24 and 38]

R. R. da Silva, P. E. Rauber, R. M. Martins, R. Minghim, and A. C. Telea. Attribute-based visual explanation of multidimensional projections. In *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2015. [page 70]

G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie. Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, 11(3):259–285, 1998. [page 41]

P. DeCamp, A. Frid-Jimenez, J. Guiness, and D. Roy. GistIcons: Seeing meaning in large bodies of literature. In *Symposium on Informaion Visualization (InfoVis) Posters*. IEEE, 2005. [page 41]

H. Dolfing. A visual analytics framework for feature and classifier engineering. Master's thesis, Department of Computer and Information Science, University of Konstanz, Konstanz, Germany, 2007. [page 126]

M. Dörk, S. Carpendale, C. Collins, and C. Williamson. Visgets: Coordinated visualizations for web-based information exploration and discovery. *Transactions on Visualization and Computer Graphics*, 14(6):1205–1212, 2008. doi: 10.1109/TVCG.2008.175. [pages 48 and 52]

M. Dörk, N. H. Riche, G. Ramos, and S. Dumais. Pivotpaths: Strolling through faceted information spaces. *Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 2012. doi: 10.1109/TVCG.2012.252. [pages 43 and 140]

W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 231–240. IEEE, 2011. [page 41]

W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually exploring large text collections using topic hierarchies. *Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013. doi: 10.1109/TVCG.2013.162. [pages 24, 38, 42, 165, 171, and 172]

C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, Dec 2012. doi: 10.1002/asi.22652. [pages 43 and 140]

E. Eaton, G. Holness, and D. McFarlane. Interactive learning using manifold geometry. In *Fall Symposium on Manifold Learning and Its Applications*, pages 10–17. AAAI, 2009. [page 99]

K. El-Arini and C. Guestrin. Beyond keyword search: Discovering relevant scientific literature. In *Knowledge Discovery in Databases (KDD)*, pages 439–447, New York, NY, USA, 2011. ACM. doi: 10.1145/2020408.2020479. [page 39]

M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim. Contovi: Multi-party conversation exploration using topic-space views. *Computer Graphics Forum*, 35 (3):431–440, 2016. [page 39]

A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 121–130. IEEE, Oct 2011. doi: 10.1109/VAST.2011.6102449. [pages 38, 70, 98, and 166]

A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Conference on Human Factors in Computing Systems (CHI)*, pages 473–482, New York, NY, USA, 2012. ACM. doi: 10.1145/2207676.2207741. [pages 38, 70, 166, 171, 173, and 174]

A. Endert, R. Burtner, N. Cramer, R. Perko, S. Hampton, and K. Cook. Typograph: Multiscale spatial exploration of text documents. In *International Conference on Big Data*, pages 17–24. IEEE, Oct 2013. doi: 10.1109/BigData.2013.6691709. [page 70]

J. A. Fails and D. R. Olsen, Jr. Interactive machine learning. In *International Conference on Intelligent User Interfaces (IUI)*, pages 39–45, 2003. doi: 10.1145/604045.604056. [page 99]

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9: 1871–1874, 2008. [page 128]

I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek. On using class-labels in evaluation of clusterings. In *KDD Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust)*, 2010. [page 30]

U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, 1996. [page 12]

P. Federico, F. Heimerl, S. Koch, and S. Miksch. A survey on visual approaches for analyzing scientific literature and patents. *Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. doi: 10.1109/TVCG.2016.2610422. [pages 7, 38, 161, 162, and 170]

J. Feinberg. Wordle. In J. Steele and N. Iliinsky, editors, *Beautiful Visualization*, pages 37–58. O'Reilly, 2010. [page 46]

J.-D. Fekete and C. Plaisant. Excentric labeling: Dynamic neighborhood labeling for data visualization. In *Conference on Human Factors in Computing Systems (CHI)*, pages 512–519, New York, NY, USA, 1999. ACM. [page 146]

J.-D. Fekete, J. J. Van Wijk, J. T. Stasko, and C. North. The value of information visualization. *Information Visualization*, pages 1–18, 2008. [page 13]

A. A. Ferreira, M. A. Gonçalves, and A. H. Laender. A brief survey of automatic methods for author name disambiguation. *SIGMOD Record*, 41(2):15–26, Aug. 2012. doi: 10.1145/2350036.2350040. [page 150]

J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 363–370, 2005. doi: 10.3115/1219840.1219885. [page 51]

J. Fogarty, D. Tan, A. Kapoor, and S. Winder. Cueflik: interactive concept learning in image search. In *Conference on Human Factors in Computing Systems (CHI)*, pages 29–38. ACM, 2008. doi: 10.1145/1357054.1357061. [page 99]

D. Fried and S. G. Kobourov. Maps of computer science. In *Pacific Visualization Symposium (PacificVis)*, pages 113–120. IEEE, 2014. doi: 10.1109/PacificVis.2014.47. [pages 42, 171, and 172]

K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda. Topigraphy: visualization for large-scale tag clouds. In *International Conference on the World Wide Web (WWW)*, pages 1087–1088. ACM, 2008. doi: 10.1145/1367497.1367669. [page 48]

G. W. Furnas. Generalized fisheye views. In *Conference on Human Factors in Computing Systems (CHI)*, pages 16–23, New York, NY, USA, 1986. ACM. doi: 10.1145/22627.22342. [page 71]

P. Gambette and J. Véronis. Visualising a text with a tree cloud. In *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V.*, pages 561–569. Springer, Berlin, Heidelberg, 2010. ISBN 978-3-642-10745-0. doi: 10.1007/978-3-642-10745-0_61. [page 49]

E. Garfield. Research fronts. *Current Contents*, Oct 1994. [page 39]

M. Gleicher. Explainers: Expert explorations with crafted projections. *Transactions on Visualization and Computer Graphics*, 19(12):2042–2051, Dec 2013. ISSN 1077-2626. doi: 10.1109/TVCG.2013.157. [page 99]

C. Goddard. *Semantic Analysis: A Practical Introduction*. Oxford Textbooks in Linguistics. Oxford University Press, 2011. ISBN 9780199560288. [page 168]

C. Görg, H. Tipney, K. Verspoor, W. A. Baumgartner Jr, K. B. Cohen, J. Stasko, and L. E. Hunter. Visualization and language processing for supporting analysis across the biomedical literature. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 420–429. Springer, 2010. [page 40]

C. Görg, Z. Liu, J. Kihm, J. Choo, H. Park, and J. Stasko. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw. *Transactions on Visualization and Computer Graphics*, 19(10):1646–1663, 2013. [pages 40 and 140]

B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. TopicNets: Visual analysis of large text corpora with topic modeling. *Transactions on Intelligent Systems and Technology*, 3(2):23:1–23:26, Feb. 2012. [page 42]

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research (JMLR)*, 3:1157–1182, 2003. [page 124]

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Exploration Newsletter*, 11(1):10–18, 2009. [page 126]

P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 2011. [page 30]

M. J. Halvey and M. T. Keane. An assessment of tag presentation techniques. In *International Conference on the World Wide Web (WWW)*, pages 1313–1314. ACM, 2007. doi: 10.1145/1242572.1242826. [page 48]

A. L. Han and D. F. Wong. Machine translation evaluation: A survey. *Computing Research Repository (CoRR)*, abs/1605.04515, 2016. [page 30]

B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378. Association for Computational Linguistics (ACL), 2011. [page 1]

M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal*, 40(1):27–37, 2003. [page 145]

S. G. Hart and L. E. Stavenland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, chapter 7. Elsevier, 1988. [page 115]

Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies (InScit)*, 2006. [page 48]

S. Havre, E. Hetzler, K. Perrine, E. Jurrus, and N. Miller. Interactive visualization of multiple query results. In *Symposium Information Visualization (InfoVis)*, pages 105–112. IEEE, 2001. [page 40]

S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: Visualizing thematic changes in large document collections. *Transactions on Visualization and Computer Graphics*, 8(1):9–20, 2002. [pages 38, 140, and 141]

M. A. Hearst. TileBars: visualization of term distribution information in full text information access. In *Conference on Human Factors in Computing Systems (CHI)*, pages 59–66, New York, NY, USA, 1995. ACM. doi: 10.1145/223904.223912. [page 39]

M. A. Hearst. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition, 2009. ISBN 0521113792, 9780521113793. [pages 15 and 68]

J. Heer and M. Agrawala. Design considerations for collaborative visual analytics. *Information Visualization*, 7(1):49–62, 2008. [page 20]

J. Heer, S. K. Card, and J. Landay. Prefuse: A toolkit for interactive information visualization. In *Conference on Human Factors in Computing Systems (CHI)*, pages 421–430. ACM, 2005. [pages 14 and 81]

F. Heimerl, C. Jochim, S. Koch, and T. Ertl. Featureforge: A novel tool for visually supported feature engineering and corpus revision. In *International Conference on Computational Linguistics (COLING) Posters*, pages 461–470, 2012a.

F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *Transactions on Visualization and Computer Graphics*, 18 (12):2839–2848, 2012b. doi: 10.1109/TVCG.2012.277. [pages 68, 71, 111, and 112]

F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word Cloud Explorer: Text analytics based on word clouds. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1833–1842, Jan 2014. doi: 10.1109/HICSS.2014.231. [pages 57 and 71]

F. Heimerl, Q. Han, S. Koch, and T. Ertl. CiteRivers: Visual analysis of citation patterns. *Transactions on Visualization and Computer Graphics*, 22(1):190–199, January 2016a. doi: 10.1109/TVCG.2015.2467621. [pages 13 and 158]

F. Heimerl, M. John, Q. Han, S. Koch, and T. Ertl. Docucompass: Effective exploration of document landscapes. In *Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Oct 2016b.

N. Heulot, M. Aupetit, and J.-D. Fekete. ProxiLens: Interactive exploration of high-dimensional data using projections. In M. Aupetit and L. van der Maaten, editors, *EuroVis Workshop on Visual Analytics using Multidimensional Projections*. The Eurographics Association, 2013. doi: 10.2312/PE.VAMP.VAMP2013.011-015. [page 72]

J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005. [pages 39 and 148]

Y. C. Ho. Abduction? Deduction? Induction? Is there a logic of exploratory data analysis? In *Annual Meeting of the American Educational Research Association*, New Orleans, LA, April 1994. [page 23]

B. Höferlin, R. Netzel, M. Höferlin, D. Weiskopf, and G. Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2012. [page 99]

T. Honkela, J. Laaksonen, H. Törrö, and J. Tenhunen. Media map: A multilingual document map with a design interface. In *Advances in Self-Organizing Maps*, pages 247–256. Springer, 2011. [page 43]

P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Moller, and J. T. Stasko. vispubdata.org: A metadata collection about ieee visualization (vis) publications. *Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. [pages 8, 52, 82, and 87]

X. Jiang and J. Zhang. A text visualization method for cross-domain research topic mining. *Journal of Visualization*, pages 1–16, 2016. doi: 10.1007/s12650-015-0323-9. [page 42]

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142. Springer, 1998. [page 33]

T. Joachims. Transductive inference for text classification using support vector machines. In *International Conference on Machine Learning (ICML)*, pages 200–209. ACM, 1999. [page 33]

C. Jochim and H. Schütze. Towards a generic and flexible citation classifier based on a faceted classification scheme. In *International Conference on Computational Linguistics (COLING)*, pages 1343–1358, 2012. [page 131]

M. John, F. Heimerl, A. Müller, and S. Koch. A visual focus+context approach for text comparison tasks. In *LREC VisLR Workshop*, 2014. [pages 7 and 10]

M. John, S. Koch, F. Heimerl, A. Müller, T. Ertl, and J. Kuhn. Interactive visual analysis of german poetics. In *Internationa Conference on Digital Humanities (DH)*. Alliance of Digital Humanities Organizations, 2015. [page 7]

P. N. Johnson-Laird. Deductive reasoning. *Annual Review of Psychology*, 50(1): 109–135, 1999. [page 23]

K. S. Jones. Natural language processing: A historical review. In *Current Issues in Computational Linguistics: In Honour of Don Walker*, pages 3–16. Springer Netherlands, 1994. ISBN 978-0-585-35958-8. doi: 10.1007/978-0-585-35958-8_1. [page 24]

A. Judea, H. Schütze, and S. Brügmann. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *International Conference on Computational Linguistics (COLING)*, pages 290–300, 2014. [page 79]

D. Jurafsky and J. H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009. ISBN 0131873210. [pages 24, 167, and 169]

I. Jusufi, A. Kerren, J. Liu, and B. Zimmer. Visual exploration of relationships between document clusters. In *International Conference on Information Visualization Theory Application (IVAPP)*, pages 195–203, 2014. [page 43]

E. Kandogan. Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 73–82. IEEE, Oct 2012. doi: 10.1109/VAST.2012.6400487. [page 70]

O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization. In *WWW Workshop on Tagging and Metadata for Social Information Organization*, 2007. [page 48]

D. Kauchak and F. Chen. Feature-based segmentation of narrative documents. In *ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing (FeatureEng)*, pages 32–39, 2005. [page 125]

W. Ke, K. Börner, and L. Viswanath. Major information visualization authors, papers and topics in the ACM library. In *Symposium on Information Visualization (InfoVis)*. IEEE, 2004. [page 139]

D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual Analytics: Scope and challenges. In *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, chapter 6, pages 76–90. Springer, 2008. [pages vi, 2, 9, 11, 12, and 26]

D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering The Information Age: Solving Problems with Visual Analytics*. Eurographics Association, 2010. ISBN 978-3-905673-77-7. [pages 2 and 9]

D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *International Conference on Information Visualisation (IV)*, pages 9–16. IEEE, 2006. [page 9]

W. Kienreich and C. Seifert. Visual exploration of feature-class matrices for classification problems. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pages 37–41, 2012. [page 125]

M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. Topiclens: Efficient multilevel visual topic exploration of large-scale document collections. *Transactions*

on Visualization and Computer Graphics, PP(99):1–1, 2016. doi: 10.1109/TVCG.2016.2598445. [pages 171 and 173]

D. Kirsh and P. Maglio. On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4):513–549, 1994. doi: 10.1207/s15516709cog1804_1. [page 11]

G. Klein, B. Moon, and R. R. Hoffman. Making sense of sensemaking 2: A macrocognitive model. *Intelligent Systems*, 21(5):88–92, 2006a. [page 20]

G. Klein, B. M. Moon, and R. R. Hoffman. Making sense of sensemaking 1: Alternative perspectives. *Intelligent Systems*, 21(4):70–73, 2006b. [pages 20 and 21]

G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso. A data-frame theory of sensemaking. In *Expertise out of context: International Conference on Naturalistic Decision Making*, pages 113–155. New York, NY, USA: Lawrence Erlbaum, 2007. [pages vi, 20, 21, 22, 23, 117, and 118]

H. Kobdani, H. Schütze, A. Burkovski, W. Kessler, and G. Heidemann. Relational feature engineering of natural language processing. In *International Conference on Information and Knowledge Management (CIKM)*, pages 1705–1708. ACM, 2010. [pages 125 and 135]

S. Koch. *Visual search and analysis of documents in the intellectual property domain*. PhD thesis, University of Stuttgart, 2012. [page 15]

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *Transactions on Visualization and Computer Graphics*, 17(5):557–569, May 2011a. doi: 10.1109/TVCG.2010.85. [pages 10 and 165]

S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *Transactions on Visualization and Computer Graphics*, 17(5):557–569, 2011b. doi: 10.1109/TVCG.2010.85. [pages 39 and 48]

S. Koch, F. Heimerl, and T. Ertl. Visual document retrieval: Supporting text search and analysis with visual analytics. *Computing in Science & Engineering*, 15(4):66–74, 2013. doi: 10.1109/MCSE.2013.93.

S. Koch, M. John, M. Wörner, A. Müller, and T. Ertl. VarifocalReader – in-depth visual analysis of large text documents. *Transactions on Visualization and Computer Graphics*, 20(12):1723–1732, Dec 2014. doi: 10.1109/TVCG.2014.2346677. [pages 10 and 39]

M. Köper, M. Zaiß, Q. Han, S. Koch, and S. Schulte im Walde. Visualisation and exploration of high-dimensional distributional features in lexical semantic classification. In *International Conference on Language Resources and Evaluation (LREC)*, pages 1202–1206, Portoroz, Slovenia, 2016. European Language Resources Association. [page 174]

R. R. Korfhage. To see, or not to see – Is that the query? In *Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 134–141, New York, NY, USA, 1991. ACM. doi: 10.1145/122860.122873. [page 40]

R. Krüger, H. Bosch, D. Thom, E. Püttmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl. Prolix - visual prediction analysis for box office success. In *Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2013. [page 7]

R. Krüger, F. Heimerl, Q. Han, K. Kurzhals, S. Koch, and T. Ertl. Visual analysis of visitor behavior for indoor event management. In *Hawaii International Conference on System Sciences (HICSS)*, Jan 2015. doi: 10.1109/HICSS.2015.139. [page 7]

B. Y.-L. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson. Tag clouds for summarizing web search results. In *International Conference on the World Wide Web (WWW)*, pages 1203–1204. ACM, 2007. doi: 10.1145/1242572.1242766. [pages 46 and 48]

K. Kurzhals, F. Heimerl, and D. Weiskopf. Iseecube - visual analysis of gaze data for video. In *Symposium on Eye Tracking Research and Applications (ETRA)*, pages 43–50. ACM, 2014. doi: 10.1145/2578153.2578158. [page 7]

K. Kurzhals, M. Hlawatsch, F. Heimerl, M. Burch, T. Ertl, and D. Weiskopf. Gaze stripes: Image-based visualization of eye tracking data. *Transactions on Visualization and Computer Graphics*, 22(1), 2015. [page 7]

K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf. Visual movie analytics. *Transactions on Multimedia*, PP(99):1–1, 2016. doi: 10.1109/TMM.2016.2614184. [pages 7 and 169]

K. Lang. Newsweeder: Learning to filter netnews. In *International Conference on Machine Learning (ICML)*, pages 331–339. ACM, 1995. [page 112]

J. H. Larkin and H. A. Simon. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1):65–100, 1987. [page 13]

Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, pages 1188–1196. ACM, 2014. [page 81]

B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. Understanding research trends in conferences using PaperLens. In *Extended Abstracts on Human Factors in Computing Systems (CHI EA)*, pages 1969–1972, New York, NY, USA, 2005. ACM. doi: 10.1145/1056808.1057069. [pages 41 and 139]

B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. SparkClouds: Visualizing trends in tag clouds. *Transactions on Visualization and Computer Graphics*, 16 (6):1182–1189, 2010. doi: 10.1109/TVCG.2010.194. [page 49]

H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 2012. doi: 10.1111/j.1467-8659.2012.03108.x. [page 42]

V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707, 1966. [page 151]

S. Levinson. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983. ISBN 9780521294140. [page 169]

D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3–12, 1994. [page 34]

D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research (JMLR)*, 5:361–397, 2004. [pages 55 and 87]

M. Ley. The DBLP computer science bibliography: Evolution, research issues, perspectives. In A. Laender and A. Oliveira, editors, *String Processing and Information Retrieval*, volume 2476 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 2002. doi: 10.1007/3-540-45735-6_1. [pages 143 and 149]

C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764 – 766, 2013. doi: 10.1016/j.jesp.2013.03.013. [page 82]

H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *Transactions on Knowledge and Data Engineering*, 17(4):491–502, april 2005. [page 126]

M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *Computing Research Repository (CoRR)*, abs/1604.07043, 2016. [page 169]

S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *Transactions on Intelligent Systems and Technology*, 3(2):25:1–25:28, 2012. doi: 10.1145/2089094.2089101. [pages 24 and 38]

S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. Topicpanorama: A full picture of relevant topics. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 183–192. IEEE, Oct 2014. doi: 10.1109/VAST.2014.7042494. [page 38]

S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker. Exploring topical lead-lag across corpora. *Transactions on Knowledge and Data Engineering*, 27 (1):115–129, Jan 2015a. doi: 10.1109/TKDE.2014.2324581. [page 43]

X. Liu, H. Shen, and Y. Hu. Supporting multifaceted viewing of word clouds with focus+context display. *Information Visualization*, 14(2):168–180, 2015b. [page 71]

Y. Liu and G. Salvendy. Design and evaluation of visualization support to facilitate decision trees classification. *International Journal of Human-Computer Studies*, 65:95–110, 2007. doi: 10.1016/j.ijhcs.2006.07.005. [page 99]

Z. Liu and J. Stasko. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *Transactions on Visualization and Computer Graphics*, 16(6):999–1008, Nov 2010. doi: 10.1109/TVCG.2010.177. [page 13]

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. J. C. H. Watkins. Text classification using string kernels. *Journal of Machine Learning Research (JMLR)*, 2:419–444, 2002. [page 120]

S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *International Conference on Human-Computer Interaction (INTERACT)*, pages 392–404. Springer, 2009. doi: 10.1007/978-3-642-03655-2_43. [pages 48 and 51]

S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *International Working Conference on Advanced Visual Interfaces (AVI)*, pages 753–756. ACM, 2012. doi: 10.1145/2254556.2254701. [pages 48, 49, and 52]

S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl. ConcentriCloud: Word cloud visualization for multiple text documents. In *International Conference on Information Visualisation (IV)*, pages 114–120. IEEE, July 2015. doi: 10.1109/iV.2015.30. [page 60]

M. Luboschik, H. Schumann, and H. Cords. Particle-based labeling: Fast point-feature labeling without obscuring other visual features. *Transactions on Visualization and Computer Graphics*, 14(6):1237–1244, Nov 2008. doi: 10.1109/TVCG.2008.152. [page 71]

A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 181–190. IEEE, Oct 2011. doi: 10.1109/VAST.2011.6102456. [pages 69, 171, and 172]

R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, M. Wade, and D. S. Ebert. Understanding syndromic hotspots - a visual analytics approach. In *Symposium on Visual Analytics Science and Technology (VAST)*, pages 35–42. IEEE, Oct 2008. doi: 10.1109/VAST.2008.4677354. [page 11]

J. D. Mackinlay, G. G. Robertson, and S. K. Card. The perspective wall: Detail and context smoothly integrated. In *Conference on Human Factors in Computing Systems (CHI)*, pages 173–176, New York, NY, USA, 1991. ACM. doi: 10.1145/108844.108870. [page 71]

K. K. Mane and K. Börner. Mapping topics and topic bursts in PNAS. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5287–5290, 2004. [page 42]

C. Manning and D. Klein. Optimization, maxent models, and conditional estimation without magic. Tutorial at HLT-NAACL 2003 and ACL 2003, 2003. [page 127]

C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999. ISBN 0-262-13360-1. [pages 30 and 31]

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715. [pages 29, 30, 33, 79, 96, and 129]

C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. [pages 50, 62, 81, 148, and 167]

R. May, P. Hanrahan, D. A. Keim, B. Shneiderman, and S. Card. The state of visual analytics: views on what visual analytics is and where it is going. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 257–259. IEEE, 2010. [page 13]

T. May and J. Kohlhammer. Towards closing the analysis gap: Visual generation of decision supporting schemes from raw data. *Computer Graphics Forum*, 27 (3):911–918, 2008. [page 99]

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Computing Research Repository (CoRR)*, abs/1301.3781, 2013. [page 81]

T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072. [pages 27 and 31]

J. Moehrmann and G. Heidemann. Efficient annotation of image data sets for computer vision applications. In *International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications (VIGTA)*, pages 2:1–2:6. ACM, 2012. doi: 10.1145/2304496.2304498. [page 98]

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2012. ISBN 978-0-470-54281-1. [page 30]

M. J. Moravcsik and P. Murugesan. Some results on the function and quality of citations. *Social Studies of Science*, 5:86–92, 1975. [page 131]

A. Moschitti, B. Coppola, D. Pighin, and R. Basili. Engineering of syntactic features for shallow semantic parsing. In *ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing (FeatureEng)*, pages 48–56, 2005. [page 125]

T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, Dec 2013. doi: 10.1109/TVCG.2013.125. [pages 29 and 99]

T. Mühlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit. Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *Transactions on Visualization and Computer Graphics*, 20(12): 1643–1652, Dec 2014. [page 166]

R. Nakazawa, T. Itoh, and T. Saito. A visualization of research papers based on the topics and citation network. In *International Conference on Information Visualization (IV)*, pages 283–289. IEEE, July 2015. doi: 10.1109/iV.2015.58. [page 40]

A. Nenkova and K. McKeown. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer, 2012. [page 78]

M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. In *Proceedings of the National Academy of Sciences*, pages 5200–5205, 2004. [pages 149 and 150]

C. North. Toward measuring visualization insight. *Transactions on Visualization and Computer Graphics*, 26(3):6–9, May 2006. doi: 10.1109/MCG.2006.70. [pages 12, 15, and 85]

L. T. Nowell, R. K. France, D. Hix, L. S. Heath, and E. A. Fox. Visualizing search results: some alternatives to query-document similarity. In *Annual International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, 1996. [page 39]

D. Oelke, D. Spretke, A. Stoffel, and D. A. Keim. Visual readability analysis: How to make your writings easier to read. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 123–130. IEEE, 2010. [pages 38, 168, 171, and 172]

D. Oelke, H. Strobelt, C. Rohrdantz, I. Gurevych, and O. Deussen. Comparative exploration of document collections: A visual analytics approach. *Computer Graphics Forum*, 33(3):201–210, June 2014. ISSN 0167-7055. doi: 10.1111/cgf.12376. [pages 38 and 42]

K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualization of a document collection: The VIBE system. *Information Processing & Management*, 29(1):69–81, 1993. [pages 38 and 40]

F. Olsson. A literature survey of active machine learning in the context of natural language processing. Technical Report 06, Swedish Institue of Computer Science, 2009. [page 34]

F. Osborne, E. Motta, and P. Mulholland. Exploring scholarly data with rexplore. In *International Semantic Web Conference (ISWC)*, pages 460–477. Springer, 2013. [page 41]

J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim. An approach to supporting incremental visual data classification. *Transactions on Visualization and Computer Graphics*, 21(1):4–17, 2015. [page 99]

F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least Square Projection: A fast high-precision multidimensional projection technique and its application to document mapping. *Transactions on Visualization and Computer Graphics*, 14(3):564–575, May 2008. doi: 10.1109/TVCG.2007.70443. [pages 70, 81, 103, and 106]

F. V. Paulovich, F. M. B. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato. Semantic wordification of document collections. *Computer Graphics Forum*, 31 (3pt3):1145–1153, June 2012. doi: 10.1111/j.1467-8659.2012.03107.x. [page 48]

C. S. Peirce. Deduction, induction, and hypothesis. *Popular Science Monthly*, 13, August 1878. [page 23]

D. Phan, L. Xiao, R. B. Yeh, P. Hanrahan, and T. Winograd. Flow map layout. In *Symposium on Information Visualization (InfoVis)*, page 29. IEEE, 2005. [pages 140 and 143]

W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009. doi: 10.1057/ivs.2009.22. [pages 14 and 23]

P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *International Conference on Intelligence Analysis*, volume 5, pages 2–4, 2005. [pages vi, 18, 19, 96, 97, 104, 105, and 119]

C. Plaisant. The challenge of information visualization evaluation. In *International Working Conference on Advanced Visual Interfaces (AVI)*, pages 109–116, New York, NY, USA, 2004. ACM. doi: 10.1145/989863.989880. [page 85]

C. Plaisant, J.-D. Fekete, and G. Grinstein. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *Transactions on Visualization and Computer Graphics*, 14(1):120–134, Jan. 2008. doi: 10.1109/TVCG.2007.70412. [page 139]

F. Poulet. Towards effective visual data mining with cooperative approaches. In S. J. Simoff, M. H. Böhlen, and A. Mazeika, editors, *Visual Data Mining*, pages 389–406. Springer, 2008. doi: 10.1007/978-3-540-71080-6_22. [page 99]

Z. Pousman, J. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007. doi: 10.1109/TVCG.2007.70541. [page 45]

U. Rajagopalan, Balajiand Lall and D. G. Tarboton. Evaluation of kernel density estimation methods for daily precipitation resampling. *Stochastic Hydrology and Hydraulics*, 11(6):523–547, 1997. [page 30]

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. [page 30]

P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, pages 1–6, 2000. [pages 36 and 37]

W. Ribarsky, B. Fisher, and W. M. Pottenger. Science of analytical reasoning. *Information Visualization*, 8(4):254–262, Dec. 2009. doi: 10.1057/ivs.2009.28. [pages 2 and 13]

P. Riehmann, M. Potthast, B. Stein, and B. Froehlich. Visual assessment of alleged plagiarism cases. *Computer Graphics Forum*, 34(3):61–70, 2015. doi: 10.1111/cgf.12618. [page 41]

E. Ringger, editor. *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, June 2005. Association for Computational Linguistics (ACL). [page 125]

A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our head in the clouds: toward evaluation studies of tagclouds. In *Conference on Human Factors in Computing Systems (CHI)*, pages 995–998. ACM, 2007. doi: 10.1145/1240624.1240775. [page 48]

G. G. Robertson and J. D. Mackinlay. The document lens. In *Symposium on User Interface Software and Technology (UIST)*, pages 101–108, New York, NY, USA, 1993. ACM. doi: 10.1145/168642.168652. [page 71]

A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014. doi: 10.1126/science.1242072. [page 82]

T. Rose, M. Stevenson, and M. Whitehead. The Reuters corpus volume 1 - From yesterday's news to tomorrow's language resources. In *International Conference on Language Resources and Evaluation (LREC)*, pages 29–31. European Language Resources Association, 2002. [page 112]

D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Conference on Human Factors in Computing Systems (CHI)*, pages 269–276. ACM, 1993. [page 18]

I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003. [page 99]

D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014. [page 19]

D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2016. [page 19]

I. A. Sag, T. Baldwin, F. Bond, A. A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for nlp. In *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 1–15. Springer, 2002. [page 51]

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988. doi: 10.1016/0306-4573(88)90021-0. [page 28]

P. Saraiya, C. North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *Transactions on Visualization and Computer Graphics*, 12 (6):1511–1522, 2006. [page 112]

M. Scaife and Y. Rogers. External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 45(2):185–213, 1996. [page 13]

C. Schätzle and D. Sacha. Visualizing language change: Dative subjects in icelandic. In *LREC Workshop VisLR II: Visualization as added value in the development, use and evaluation of Language Resources*, 2016. [page 25]

G. Schneider, M. El-Assady, and H. M. Lehmann. Tools and methods for processing and visualizing large corpora. In *Studies in Variation, Contacts and Change in English*. VARIENG Helsinki, 2016. [page 174]

G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning (ICML)*, pages 839–846. ACM, 2000. [pages 34 and 115]

J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Conference on Human Factors in Computing Systems (CHI)*, pages 2037–2040. ACM, 2009. doi: 10.1145/1518701.1519010. [page 48]

T. Schreck and D. Keim. Visual feature space analysis for unsupervised effectiveness estimation and feature engineering. In *International Conference on Multimedia and Expo (ICME)*, pages 9–12. IEEE, 2006. [page 125]

T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193, 2010. doi: 10.1057/ivs.2010.2. [page 72]

K. K. Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2006. [page 134]

S. Scott and S. Matwin. Feature engineering for text classification. In *International Conference on Machine Learning (ICML)*, pages 379–388. ACM, 1999. [page 125]

S. D. Scott, N. Lesh, and G. W. Klau. Investigating human-computer optimization. In *Conference on Human Factors in Computing Systems (CHI)*, pages 155–162, New York, NY, USA, 2002. ACM. doi: 10.1145/503376.503405. [page 10]

J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3: 417–424, 1980. [page 2]

C. Seifert and M. Granitzer. User-based active learning. In *International Conference on Data Mining Workshops (ICDMW)*, pages 418–425. IEEE, 2010. doi: 10.1109/ICDMW.2010.181. [page 98]

C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer. On the beauty and usability of tag clouds. In *International Conference on Information Visualisation (IV)*, pages 17–25. IEEE, 2008. doi: 10.1109/IV.2008.89. [page 48]

C. Seifert, V. Sabol, and M. Granitzer. Classifier hypothesis generation using visual analysis methods. In *International Conference on Networked Digital Technologies*, pages 98–111, 2010. [page 98]

C. Seifert, J. Jurgovsky, and M. Granitzer. Facetscape: A visualization for exploring the search space. In *International Conference on Information Visualisation (IV)*, pages 94–101. IEEE, July 2014. doi: 10.1109/IV.2014.49. [page 49]

J. Seo and B. Shneiderman. Interactively exploring hierarchical clustering results. *Computer*, 35(7):80–86, 2002. [page 129]

B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. [pages 33 and 34]

B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. ISBN 978-1608457250. [page 33]

D. Shahaf, C. Guestrin, and E. Horvitz. Metro maps of science. In *Knowledge Discovery in Databases (KDD)*, pages 1122–1130, New York, NY, USA, 2012. ACM. [page 39]

B. Shaparenko, R. Caruana, J. Gehrke, and T. Joachims. Identifying temporal patterns and key players in document collections. In *ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications*, pages 165–174. IEEE, 2005. [page 148]

H. Sharara, A. Sopan, G. Namata, L. Getoor, and L. Singh. G-pare: A visual analytic tool for comparative analysis of uncertain graphs. In *Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2011. doi: 10.1109/VAST.2011.6102442. [page 43]

B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Symposium on Visual Languages*, pages 336–343. IEEE, Sep 1996. doi: 10.1109/VL.1996.545307. [pages 14 and 162]

Y. B. Shrinivasan and J. J. van Wijk. Supporting the analytical reasoning process in information visualization. In *Conference on Human Factors in Computing Systems (CHI)*, pages 1237–1246. ACM, 2008. [page 13]

J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, 2008. doi: 10.1177/0165551506078083. [pages 47 and 48]

A. Skupin. A cartographic approach to visualizing conference abstracts. *Computer Graphics and Applications*, 22(1):50–58, 2002. [page 42]

A. Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101(Suppl 1): 5274–5278, 2004. [page 42]

D. Smilkov, S. Carter, D. Sculley, F. Viegas, and M. Wattenberg. Direct-manipulation visualization of deep networks. In *SIGKDD Workshop on Interactive Data Exploration and Analytics*, pages 115–118. ACM, 2016. [page 169]

G. Smith. *Tagging: People-Powered Metadata for the Social Web*. New Riders, 2008. ISBN 978-0-321-52917-6. [page 46]

P. Somol, J. Novovičová, and P. Pudil. Efficient feature subset selection and subset size optimization. In *Pattern Recognition Recent Advances*, chapter 4, pages 75–97. InTech, 2010. [pages 124 and 126]

R. Spence and M. Apperley. Data base navigation: an office environment for the professional. *Behaviour & Information Technology*, 1(1):43–54, 1982. doi: 10.1080/01449298208914435. [page 71]

J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *Transactions on Visualization and Computer Graphics*, 22(1):629–638, 2016. doi: 10.1109/TVCG.2015.2467717. [pages 72 and 76]

J. Stasko. Value-driven evaluation of visualizations. In *Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, pages 46–53, New York, NY, USA, 2014. ACM. doi: 10.1145/2669557.2669579. [page 15]

J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008. doi: 10.1057/palgrave.ivs.9500180. [pages 38 and 48]

J. Stasko, J. Choo, Y. Han, M. Hu, H. Pileggi, R. Sadana, and C. D. Stolper. Citevis: Exploring conference paper citation data visually. In *Conference on Information Visualization Posters*. IEEE, 2013. [page 140]

M. Stefaner. Visual tools for the socio-semantic web. Master thesis, University of Applied Sciences Potsdam, 2007. [page 48]

F. Stoffel, L. Flekova, D. Oelke, I. Gurevych, and D. A. Keim. Feature-based visual exploration of text classification. In *VIS Symposium on Visualization in Data Science (VDS)*. IEEE, 2015. [page 126]

H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim, and O. Deussen. Document cards: A top trumps visualization for documents. *Transactions on Visualization and Computer Graphics*, 15(6):1145–1152, 2009. [page 40]

G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. Zhu, and R. Liang. Evoriver: Visual analysis of topic coopetition on social media. *Transactions on Visualization and Computer Graphics*, 20(12), 2014. [page 39]

D. R. Swanson. Complementary structures in disjoint science literatures. In *Annual International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 280–289, New York, NY, USA, 1991. ACM. doi: 10.1145/122860.122889. [page 3]

D. R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91 (2):183–203, Apr. 1997. doi: 10.1016/S0004-3702(97)00008-8. [page 3]

Tagul. Gorgeous tag clouds. http://tagul.com, 2009. [page 47]

Tagxedo. Word cloud with styles. http://www.tagxedo.com, 2016. [page 47]

J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *Knowledge Discovery in Databases (KDD)*, pages 990–998, New York, NY, USA, 2008. ACM. doi: 10.1145/1401890.1402008. [page 147]

S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 103–110, 2006. [page 131]

D. Thom. *Visual analytics of social media for situation awareness*. PhD thesis, University of Stuttgart, 2015. [page 23]

D. Thom and T. Ertl. Treequest: A treemap-based query sandbox for microdocument retrieval. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1714–1723, Jan 2015. doi: 10.1109/HICSS.2015.206. [page 49]

D. Thom, H. Bosch, and T. Ertl. Inverse document density: A smooth measure for location-dependent term irregularities. In *International Conference on Computational Linguistics (COLING)*, pages 2603–2618, 2012a. [pages 29 and 120]

D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Pacific Visualization Symposium (PacificVis)*, pages 41–48. IEEE, 2012b. [pages 80 and 120]

J. Thomas and J. Kielman. Challenges for visual analytics. *Information Visualization*, 8(4):309–314, 2009. doi: 10.1057/ivs.2009.26. [page 13]

J. J. Thomas and K. A. Cook. *Illuminating the path*. IEEE, 2005. ISBN 0769523234. [pages 9 and 11]

J. J. Thomas and K. A. Cook. A visual analytics agenda. *Computer Graphics and Applications*, 26(1):10–13, Jan. 2006. doi: 10.1109/MCG.2006.5. [page 11]

M. Tobiasz, P. Isenberg, and M. S. T. Carpendale. Lark: Coordinating co-located collaboration with information visualization. *Transactions on Visualization and Computer Graphics*, 15(6):1065–1072, 2009. [page 140]

C. Tominski, S. Gladisch, U. Kister, R. Dachselt, and H. Schumann. A survey on interactive lenses in visualization. In R. Borgo, R. Maciejewski, and I. Viola, editors, *EuroVis State-of-the-Art Reports*. The Eurographics Association, 2014. doi: 10.2312/eurovisstar.20141172. [page 71]

S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, 2001. [page 34]

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, pages 173–180. Association for Computational Linguistics (ACL), 2003. doi: 10.3115/1073445.1073478. [page 51]

E. R. Tufte. *Envisioning information*. Graphics Press LLC, 1990. ISBN 978-0961392116. [page 13]

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977. ISBN 978-0201076165. [pages 3 and 17]

S. van den Elzen and J. van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *Conference on Visual Analytics Science and Technology (VAST)*, pages 151–160. IEEE, 2011. doi: 10.1109/VAST.2011.6102453. [pages 99 and 166]

L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. [pages 70, 81, and 86]

F. van Ham. Case study: Visualizing visualization. In *Symposium on Information Visualization (InfoVis)*. IEEE, Oct 2004. doi: 10.1109/INFVIS.2004.14. [page 43]

F. van Ham, M. Wattenberg, and F. B. Viégas. Mapping text with Phrase Nets. *Transactions on Visualization and Computer Graphics*, 15(6):1169–1176, 2009. [pages 25, 171, 172, and 174]

J. J. Van Wijk. The value of visualization. In *Visualization*, pages 79–86. IEEE, 2005. [page 15]

V. Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4. [page 31]

F. B. Viégas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008. [pages 5 and 46]

F. B. Viégas, S. Golder, and J. Donath. Visualizing email content: Portraying relationships from conversational histories. In *Conference on Human Factors in Computing Systems (CHI)*, pages 979–988, New York, NY, USA, 2006. ACM. doi: 10.1145/1124772.1124919. [page 36]

F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: a site for visualization at internet scale. *Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007. doi: 10.1109/TVCG.2007.70577. [pages 47 and 49]

F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, 2009. doi: 10.1109/TVCG.2009.171. [page 142]

U. Von Luxburg. A tutorial on spectral clustering. *Statistical Computing*, 17(4): 395–416, 2007. [page 35]

S. Vrochidis, H. Bosch, A. Moumtzidou, F. Heimerl, T. Ertl, and Y. Kompatsiaris. An environmental search engine based on interactive visual classification. In *International Workshop on Multimedia Analysis for Ecological Data (MAED)*, pages 49–52, New York, NY, USA, 2012. ACM. doi: 10.1145/2390832.2390845. [page 7]

R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. What's being said near "Martha"? Exploring named entities in literary text collections. In *Symposium on Visual Analytics Science and Technology (VAST)*, pages 107–114. IEEE, 2009. [page 48]

J. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58:236–244, 1963. [page 129]

C. Ware. *Visual Thinking for Design*. Morgan Kaufmann, 2010. ISBN 978-0123708960. [page 13]

C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, CA, USA, 3rd edition, 2012. ISBN 9780123814647, 9780123814654. [pages 9 and 13]

M. Ware, E. Frank, G. Holmes, M. Hall, and I. H. Witten. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292, 2002. [page 99]

M. Wattenberg. Baby names, visualization, and social data analysis. In *Symposium on Information Visualization (InfoVis)*, pages 1–7. IEEE, Oct 2005. [page 140]

K. E. Weick. The collapse of sensemaking in organizations: The mann gulch disaster. *Administrative Science Quarterly*, pages 628–652, 1993. [page 20]

K. E. Weick. *Sensemaking in Organizations*. Foundations for Organizational Science. SAGE Publications, 1995. ISBN 9780803971776. [page 20]

K. E. Weick, K. M. Sutcliffe, and D. Obstfeld. Organizing and the process of sensemaking. *Organization Science*, 16(4):409–421, 2005. [page 20]

J. D. West. *Eigenfactor: ranking and mapping scientific knowledge*. PhD thesis, University of Washington, 2010. [page 140]

J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Symposium on Information Visualization (InfoVis)*, pages 51–58. IEEE, Oct 1995. doi: 10.1109/INFVIS.1995.528686. [pages 38, 68, 69, and 70]

P. C. Wong and J. Thomas. Visual analytics. *Computer Graphics and Applications*, 24(5):20–21, Sept 2004. doi: 10.1109/MCG.2004.39. [pages 2, 9, and 11]

P. C. Wong, B. Hetzler, C. Posse, M. Whiting, S. Havre, N. Cramer, A. Shah, M. Singhal, A. Turner, and J. Thomas. IN-SPIRE InfoVis 2004 contest entry. In *Symposium on Information Visualization (InfoVis)*, pages 51–52. IEEE, 2004. [pages 41 and 139]

Wordle. Beautiful word clouds. http://www.wordle.net, 2014. [page 47]

Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: Interactive visualization of hotel customer feedback. *Transactions on Visualization and Computer Graphics*, 16(6):1109–1118, 2010. doi: 10.1109/TVCG.2010.183. [page 49]

Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-preserving word clouds by seam carving. *Computer Graphics Forum*, 30(3):741–750, 2011. [pages 41 and 48]

Y. Wu, S. Liu, K. Yan, M. Liu, and F. Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *Transactions on Visualization and Computer Graphics*, 20(12):1763–1772, Dec 2014. doi: 10.1109/TVCG.2014.2346920. [pages 39 and 140]

P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu. Visual analysis of topic competition on social media. *Transactions on Visualization and Computer Graphics*, 19(12):2012–2021, 2013. doi: 10.1109/TVCG.2013.221. [page 39]

J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, page 4. ACM, 2008. [pages 15 and 18]

J. Zhang, C. Chen, and J. Li. Visualizing the intellectual structure with paper-reference matrices. *Transactions on Visualization and Computer Graphics*, 15 (6):1153–1160, Nov 2009. [page 140]

J. Zhang, B. Ahlbrand, A. Malik, J. Chae, Z. Min, S. Ko, and D. S. Ebert. A visual analytics framework for microblog data analysis at multiple scales of aggregation. *Computer Graphics Forum*, 35(3):441–450, 2016. doi: 10.1111/cgf.12920. [pages 171, 173, and 175]

J. Zhao, F. Chevalier, C. Collins, and R. Balakrishnan. Facilitating discourse analysis with interactive visualization. *Transactions on Visualization and Computer Graphics*, 18(12):2639–2648, 2012. [pages 25 and 169]

J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *Transactions on Visualization and Computer Graphics*, 19(12):2080–2089, 2013. doi: 10.1109/TVCG.2013.167. [page 69]

J. Zhao, N. Cao, Z. Wen, Y. Song, Y. R. Lin, and C. Collins. Fluxflow: Visual analysis of anomalous information spreading on social media. *Transactions on Visualization and Computer Graphics*, 20(12):1773–1782, Dec 2014. doi: 10.1109/TVCG.2014.2346922. [page 39]