DocuCompass: Effective Exploration of Document Landscapes

Florian Heimerl, Markus John, Qi Han, Steffen Koch *Member, IEEE*, and Thomas Ertl *Member, IEEE*

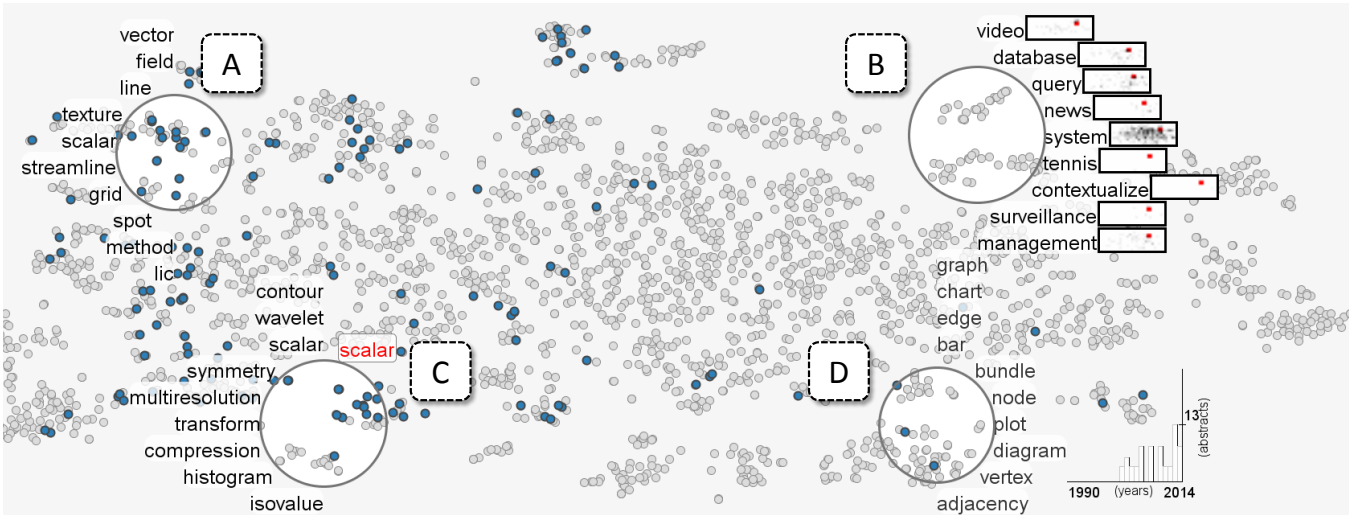

Figure 1: DocuCompass comprises lenses with different features: (A) a lens showing terms as text labels for characterizing focused documents, (B) a lens depicting previews of term distributions over the display, (C) a lens with the term 'scalar' selected, (D) a lens using a bar chart to depict the distribution of publication years.

Abstract

The creation of interactive visualization to analyze text documents has gained an impressive momentum in recent years. This is not surprising in the light of massive and still increasing amounts of available digitized texts. Websites, social media, news wire, and digital libraries are just few examples of the diverse text sources whose visual analysis and exploration offers new opportunities to effectively mine and manage the information and knowledge hidden within them. A popular visualization method for large text collections is to represent each document by a glyph in 2D space. These landscapes can be the result of optimizing pairwise distances in 2D to represent document similarities, or they are provided directly as meta data, such as geo-locations. For well-defined information needs, suitable interaction methods are available for these spatializations. However, free exploration and navigation on a level of abstraction between a labeled document spatialization and reading single documents is largely unsupported. As a result, vital foraging steps for task-tailored actions, such as selecting subgroups of documents for detailed inspection, or subsequent sense-making steps are hampered. To fill in this gap, we propose DocuCompass, a focus+context approach 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 allows for effective interactive exploration of document landscapes without disrupting the mental map of users by changing the layout itself. We discuss the suitability of multiple navigation and characterization methods for different spatializations and texts. Finally, we provide insights generated through user feedback and discuss the effectiveness of our approach.

2016 IEEE Conference on Visual Analytics Science and Technology (VAST)
23-28 October, Baltimore, Maryland, USA
978-1-5090-5661-3/16/$31.00 ©2016 IEEE

Keywords: interaction techniques, document visualization, text mining, visual analytics, focus+context

1 Introduction

Visual text analysis has attracted a lot of attention during recent years. Although this has led to many impressive analysis techniques, written text still has to be considered special compared to other data types with known structure. Written natural languages are the most universal visual systems of symbols that exists [53], and are capable of encoding very complex information. Therefore, using visualization to aggregate and analyze text is only a viable option, if the goal is to either abstract or summarize texts, or to extract and explicate very specific details. Due to the semantic gap, even these very abstract and narrow goals are only achievable if writing itself is used within visualizations, avoiding to decipher the full meaning of a text. This is most often done by selecting terms that are extracted and used as labels, as part of visual abstractions and explications. Thus users' world knowledge and language comprehension skills are used as an integral part of analysis.

Many visual methods for analyzing text are designed for information needs, analysis goals, or extraction tasks that are known in advance. These methods require users' previous knowledge about the contents of texts and the information they are looking for, i.e. a well-defined information need [21]. Although exploratory analysis methods for text datasets have been proposed, free exploration with no or very little prior knowledge about the corpus is hampered due to missing suitable interaction techniques. We aim to improve this situation with DocuCompass. It works with 2D document spatializations [54], 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. [10], it

can be considered a cue-based focus+context technique, depending on the configuration (discussed in Section 3). Through the large number of configuration possibilities, many different foraging phases [41], 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 visual analytics tasks and systems in previous work [5, 23, 4]. 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 contributions:

- The technique we present 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.
- We report and discuss 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.
- We offer an initial user study 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 text details, and lack some intermediate interaction method.

2 Related Work

DocuCompass contributes to *visual text analysis* by offering 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 for unspecified information needs.

2.1 Visual Text Analysis

Many visual approaches aim to globally summarize text collections or extract specific properties. They typically assume that information needs, analysis goals, or extraction tasks are known beforehand. Themeriver [20] is an approach to depict temporal dynamics of word occurrences. It helps understand trends and developments of theme occurrence over time. Jigsaw [49] is an approach that offers a variety of integrated visual tools to relate entities from documents. This helps recognizing and extracting complex events that are visually depicted and summarized. The approach uses named entity recognition to extract mentions of entities and coalesces them with document meta data, such as creation dates, to gather additional information about events. Parallel tag clouds [11] makes use of multiple word clouds to track term use in corpora with different facets. More recent 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) [3], has become a popular method to aggregate large document sets based on their term distribution. Two recent approaches [15, 13] allow users to visually explore a hierarchy of topics derived from a text dataset, offering drill-down interaction. Other approaches, e.g., [7] do not depict temporal dynamics of topics, but allow users to visually adjust the topic model to their analysis goals. All approaches mentioned so far either aggregate or abstract document meta data or content globally. Some of them do not contain a visual representation of single documents at the overview level. As document boundaries are natural demarcations of text within corpora, representing them as single visual objects and allowing the analysis of their content and relations can help analysts a lot in making sense of a text collection. One possibility to achieve this is depicting documents in a spatial layout which has been done in a variety of different forms.

2.2 Spatialization of Texts

A straightforward way to lay out documents in 2D is with an intrinsic 2D mapping, such as geo-locations [33]. Alternative methods have been devised either based on meta data or content. Galaxies [54] 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. Such approaches are often based on the vector space model, a common representation for documents as vectors in vocabulary space [35]. Document galaxies are created by mapping this high-dimensional space into 2D, preserving pairwise distances as well as possible to make them visually interpretable. Since then, the idea has been used often including commercial packages such as IN-SPIRE^{TM} [54] and Aureka [1]. Correll et al. [12] use it in an approach to explore collections of tagged texts.

Mapping the high-dimensional vectors into 2D introduces errors with respect to the pairwise distances. Recent approaches to reduce these errors are least squares projection (LSP) [40] and t-distributed stochastic neighbor embedding (t-SNE) [52]. 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 devised [14, 27]. Though this concept has been applied to high-dimensional text data [7, 54], a very large number of dimensions makes selecting useful areas and terms difficult. 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 users to adapt the underlying placement model. Endert et al. [18, 17] let users move documents in projection space, e.g., to display similar documents closer together. Their 'semantic interaction' concept incorporates user feedback into document placement decisions. The problem of effective exploration to understand the 2D layout, however, remains unsolved. Users still have to read single documents, or start with an initial information need formulated as a search query. Typograph [16] 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 Typograph lays out keywords extracted from documents, our lens-based technique directly operates on document layouts. 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 possible. In addition, our technique can be integrated

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

with a wide variety of document characterization methods, based on contents and meta data.

2.3 Focus+Context Interaction and Lenses

Different focus+context techniques exist for various types of data [10], including *magic lens* techniques for text analysis. Around the time Furnas [19] proposed fisheye views[2] for structured information, Spence and Apperly [47] published their approach to quickly access large amounts of text. Later, a focus+context technique was developed [34] that provides meta data context of documents and shows textual details in the focus region. The 'document lens' [44] was proposed as 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 [51] are a versatile [2] and straightforward way to realize the focus+context concept. Only few magic lens approaches exist to explore and navigate in text collections. To the best of our knowledge, they made their first appearance in our contribution to VAST challenge 2011 to explore geo-located micro blog messages [5]. We extended their use to other domains, such as disaster management based on micro blog messages [4], and the exploration of larger documents [23]. These approaches are mostly limited to document frequency to weigh and select terms that provide content information about focused documents.

We aim to improve magic lens approaches for text analysis with regard to multiple aspects. One is to support analytic tasks as much as possible. The lens technique is particularly useful during exploration, or the 'foraging' stage [41]. Analysis scenarios in which users start out with an unspecified or very coarse information need [21] 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 problem in itself [32]. Word clouds have also been extended in multiple directions into interactive visual interfaces for text analysis [31, 24]. 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 separated 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. [1] discuss placement variants in the context of depicting names next to focused regions, including cues for linking individual names and objects. 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 data [9] and high-dimensional data in general [46]. Stahnke et al. [48] focus on understanding 2D layouts after dimensionality reduction and the contribution 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 information loss [25]. As opposed to these approaches, our techniques in particular focuses on exploring and characterizing text documents. We also discuss methods that can be integrated with our lens-based technique to reduce local ambiguity and uncertainties caused by information loss.

3 The DocuCompass Technique

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: **1)** Facilitate flexible analysis on arbitrary levels of granularity of the dataset; **2)** Provide characterizations of the focused document set that can be computed at interactive rates and that can be easily followed and quickly processed by humans; **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 datasets. In the remainder of this section, we discuss the design of DocuCompass based on these five goals.

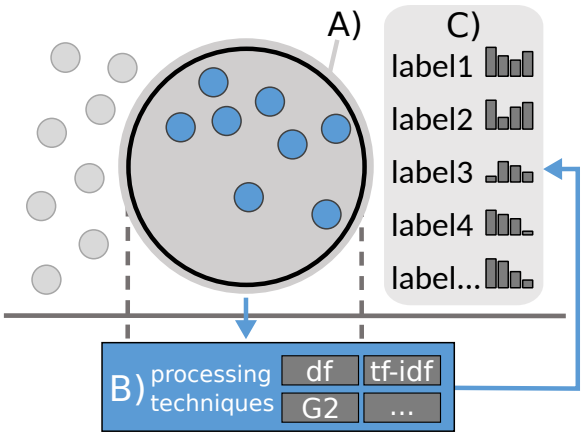

Figure 2: Overview of the DocuCompass technique. It includes three tightly integrated components for interactive exploration of 2D landscapes of text documents. The first one is a flexible visual metaphor to help users specify a set of interesting documents (A). Once the set of documents are chosen, they can be further processed by different text processing techniques to extract and sort terms for characterizing the documents (B). A third visual component identified by users or the system can then be updated to reflect the user's intention (C).

DocuCompass consists of two building blocks (see Figure 2). The first one comprises the lens to focus document sets in a visual spatialization, and a visual characterization of the focused texts, which is continuously updated during lens movement. The second one are a set of document analysis techniques, which feed the visual characterization with their results. DocuCompass can be flexibly configured with different document analysis techniques and a variety of visual representations. Since DocuCompass is designed as a focus+context technique, it is connected to the underlying spatialization in two ways. It is important that the spatial distribution is made available programmatically for analyzing focused documents and to show navigation cues outside the lens. All of these aspects are discussed subsequently including different combinations and configurations of DocuCompass, as well as their benefits and drawbacks based on application scenarios.

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. This can be seen in Figure 1(A). This lets analysts quickly explore certain regions of a 2D document landscape and get an idea of the contents and the diversity of focused documents. In partial fulfillment of **goal 1**, we have decided to show roughly ten terms. This keeps helps users to maintain an

[2] A not officially published version has been available since 1981, c.f. [6]

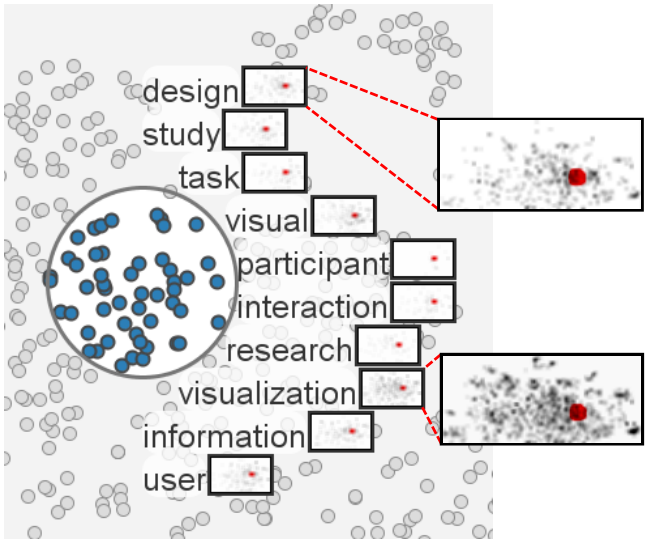

Figure 3: A lens with mini heat maps offers a preview of term distributions. As the differences between heat maps come across as rather subtle in the screen shot, we have enlarged two of them for illustration. The term "visualization" is much more prominent in the dataset than the term "design", although 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.

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 besides 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, as is the case with overview and detail approaches. Bertini et al. [1] 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 means, however, that occlusion concerns document close to the lens, that users might be interested in (cf. **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 increased 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 analysts 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 edges otherwise.

Global Navigation In partial fulfillment of **goal 4**, we have devised methods that help users with globally navigating the document scatterplot. Previews of a term's distribution can be displayed optionally as small multiples of the spatialization. They include a heat map that gives an overview of how frequent a term is used in other documents, or other visual representations. This helps users to assess quickly which term might be of interest to them. Figure 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 we often encountered situations in which exploration manifests a more concrete information need, e.g., by bringing up an interesting term. The highlighting of respective document glyphs is done with preattentively perceptible encoding (here color) in order to reduce the time required for planning subsequent exploration tasks. This is shown in Figure 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 (cf. **goal 3**).

Exploring and analyzing a 2D document landscape is a challenging and complex task for analysts. In layouts generated with dimensionality reduction methods, exploring the space involves discovering different topics and uncovering the general structure of the documents space. Exploring geographical layouts involves finding and analyzing interesting regions and learn 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.

Local Navigation The second requirement of **goal 4** is local navigation. Previously discussed techniques are useful for local placement of the lens, including term distributions as heat maps (Figure 3) or highlights when terms are moused over. However, as the adaption of local lens placement can be quite intricate, we include a number of additional local techniques that support 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. [48] 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 by clustering focused documents based on their high-dimensional representations. Once analysts activate this functionality, clusters are displayed by coloring document glyphs (see Figure 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 gives analysts information about the simi-

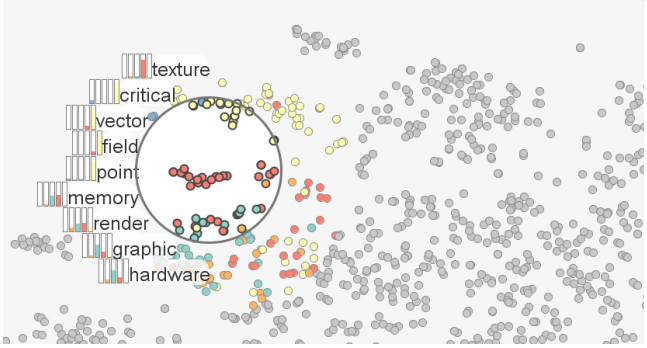

Figure 4: A lens depicts clusters of documents by coloring document glyphs accordingly. Bar charts with respective colors are placed near the terms to indicate the relative prominence of the terms in different clusters. If users are interested in topics like "vector field" and "critical point", they might want to move the lens up to include more related documents.

larity structure of the focused documents and helps with navigating locally. By learning about which of the documents are particularly close, and which are less close in the original space, they 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 Fig. 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. Analysts are thus supported with optimizing the placement of the lens once they have discovered an interesting region in the 2D document space. 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. Analysts can adjust their exploration strategy to a more fine-grained approach in the former case, and a more coarse grained approach for regions with the latter structure.

3.2 Document Characterization

Many methods have been proposed in natural language processing to summarize documents [38]. 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 analysts, it is important to get an overview of the main contents of a document set. This helps them to develop and solidify an information need for the corpus, and 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. They thus allow for flexible and comprehensive analyses. 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 count the number of documents that a term occurs in. It is called *Document frequency* (df). 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 to rate terms with the *term frequency - inverse document frequency* (tf-idf) method. It uses the logarithm of a term's inverse df to counterbalance its term frequency within a document. Thus, less emphasis is put on terms popular throughout the dataset. Common terms that have little discriminating information are thus effectively removed from the list. G^2 is another rating scheme [43]. It compares the frequency of terms in the documents under the lens to those in the remaining dataset [11]. Terms that have a much higher frequency in the focused documents are rated higher than others. The measure thus selects the terms that help to distinguish between the focused documents and all others. Chuang et al. [8] find that G^2 reaches a high keyword extraction accuracy compared to other measures based on keywords selected by human users. Especially for uniform datasets 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.

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 the use of a stop word lists. It contains terms that do not contain information in isolation, and are thus removed from the term selection altogether. Using a stop word list thus eliminates highly frequent terms, such as pronouns, conjunctions, and prepositions. A number of additional methods can be used for term selection or refinement of selections. Stemming and lemmatization [35], for example, are popular methods when extracting terms and creating word vectors. 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* $\rightarrow$ *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. There is a number of additional methods that have proven helpful to characterize documents. One that helps to filter terms of different types is part-of-speech (POS) tagging. It automatically finds the part-of-speech of each token in a text. If POS information is available, users can filter, e.g., by nouns, verbs, or adjectives. Depending on the types of text, this can help to reduce noise and filter out non-relevant information [24]. A method similar to POS tagging is named entity recognition (NER). It classifies all nouns within a text according to whether or not they are part of a named entity, and outputs sequences that constitute proper names.

Each of them is also classified into different types, including persons, locations, organizations. Filtering by them can provide additional information about important aspects of a document set. For example, for literary texts, frequently occurring characters might be relevant, while for newspaper articles, users may filter for mentioned locations. Further methods exist that might be useful for specific types of documents or analysis scenarios, such as extracting and filtering technical vocabulary [26] for patent documents.

Metadata Another way of characterizing focused documents is through meta data. The types available are dependent on the documents. Analyses of narrative texts can be enriched 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 or linked persons, or the number of retweets are potentially important information. Patent datasets, e.g., include applications, filing dates, and citations that could be summarized for a focused set of patents. For scientific literature, available meta data includes authors, their affiliations, the conference or the journal of a publication, the citations, and its index terms. Meta data 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 (cf. Figure 1D), or number of citations over time.

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 coordinates 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 contain the location of their applicant, or hotel reviews with the location of the reviewed hotel. Using these geo-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 [50]. These terms are extracted and displayed on the map and help users with global navigation. 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 analysts can explore using different types of lenses. 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 meta data or textual content. In the former case, the simplest way to map documents into 2D is to select two meta data 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 the number of characters they contain, and their number of retweets. 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 complexities. 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 [40] and tSNE [52]. Text documents are typically represented by the vector space model. A document vector contains a term distribution using either the raw frequency or any of the measures discussed in Section 3.2. Comparing these distributions gives a good account of their content similarity. Recently, methods to statistically learn term similarities have been proposed [37] and shown to achieve high document classification accuracy [28]. This suggests that they might be useful for creating high-accuracy document spatializations as well. The axes of the resulting 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 Prototypical 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 [22]. All natural language processing, including tokenization, sentence splitting, and lemmatization is done by Stanford CoreNLP [36]. In addition, the prototype includes content-based spatialization methods for documents, namely LSP [40] or t-SNE [52], based on their respective libraries. The prototype can be pointed to a text dataset 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 characterization methods can be activated through a context menu on the plot. 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 plot that shows publication dates of documents over time (cf. Figure 1(D)), if available in the dataset. In addition, heat map previews are available for each of the keyword measures. Hovering and selecting terms works as previously described in Section 3. Finally, for local navigation, the previously described clustering methods has been implemented (cf. Figure 4). Technically, we use a modified version of the clustering algorithm proposed by Rodriguez et al. [45]. It 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 scheme 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 very high local density, and high distances to other centroid candidates based on the median absolute deviation method [30]. Automatically estimating the number of clusters frees users from the burden of having to provide any parameters for clustering during exploration.

5 User Feedback

Generating insights into our interaction technique through user feedback is difficult. Beyond the typical problems that make visualization and interaction evaluation challenging [42], 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, even if some of the mentioned facets were fixed. Also, different corpora would have to be used to rule out learning effects. Testing explorative interaction techniques designed for underspecified information needs is particularly challenging. This is due to the fact that it is not possible to clearly define what successful exploration means. 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 [39] did seem adequate with such open-ended exploration tasks as the insights and their complex interplay in generating an information need 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 their feedback. To compare a classical exploration approach to our new technique, we restricted many of the facets described above. Such a setup certainly has problems, e.g., that static labeling already helps a lot with the interpretability of the used document projections. However, if we would have included such a static labeling, it would have been very difficult to attribute findings to either the static labels or the DocuCompass technique. Despite these problems, we find that the evaluation suggests the adequacy of our approach. In addition, a number of advantages as opposed to classical interaction techniques can be elicited from our test users' feedback.

5.1 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 dataset, 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 [52] to project all of the datasets 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 the upper one contains its full text, the lower one characterizes the current document with a word cloud. Users can choose between tf-idf and G^2 for term selection methods. 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. Users are able to activate several different radial 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.

5.2 Participants and Procedure

Overall we had nine participants, one female and eight males. Their average age was 31 years (min 27, max 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 worked in the field of natural language processing (NLP), two as post-doctoral researchers, and one as a PhD candidate. All of them had basic knowledge in information visualization and one of them had heard about lens techniques. 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 datasets for the study. The first one was used for an introductory session and contained all paper abstracts of VIS publications [3]. The second and third dataset were Reuters news wire texts selected according to their topic (sports news, and international conflicts) from the Reuter's RCV1 corpus [29]. We permuted each of the four possible combinations of implementation and dataset, asking each participant to complete an exploration task with both implementations on different datasets. For each participant, we also changed the order of the implementations, alternating between the simple and the lens-based implementation as the first one to work with. For both datasets, we posed the same three questions to each of the participants to answer based on the results of their explorations. The first questions asked about the general theme of the entire dataset, the second one asked about topics that users could identify within the spatialization, and the third one asked for any subtopics into which identified topics could be split up. 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 for 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 dataset. **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 potential extensions to the basic approach that we implemented using additional questions. We conducted the study session according to the think-aloud paradigm, asking participants to voice any of their thoughts during the entire sessions and recording everything on paper.

5.3 Results

In the following, we report the results of the user feedback sessions, and discuss whether and why they were to be expected. 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 dataset, then shrink the lens to focus

[3] We used the dataset available at vispubdata.org

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 dataset, 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 counterbalanced by an often general reluctance of participants to read complete texts.

As we expected, the NLP researchers all found that hovering over a term 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 them 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 mentioned the possibility of using multiple lenses at once as being very beneficial for analysis. Further, one visualization researcher positively mentioned 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 mentally process 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 great looking tool for exploration of texts that provides quick overviews of datasets. 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.

6 Discussion

This section discusses user feedback and scalability aspects. 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 switch between different visual representations of DocuCompass during the analysis (on-the-fly), instead of using a different lens. This is certainly true and we plan to implement such a possibility in the future, since this supports a more flexible analysis. Another remark was that multiple lenses might be more helpful if relations between lenses are 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 vector representations 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. For dynamic sets, updating data structures when new documents are added or older ones removed introduces additional complexity and effort. In these situations, the cost of updating the data structure continuously has to be balanced against straight-forward methods.

Information scalability is the ability to retrieve relevant information from large data collections. Depending on the exploration

task, different modes of aggregation are appropriate to achieve this. Analysts with initially underspecified information need can follow different strategies during exploration, depending on their new insights and aspects of the dataset they develop an interest for. Possible strategies include general exploration of themes, comparative analyses of different parts of the set, navigational or 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 on which technique to use or to restrict it to a specific one, depends on the task, scenario, and users' expertise. An aspects we underestimated when developing first variants of the technique are the influence of text types. As mentioned before, text types greatly influence the usefulness of certain mining techniques.

Visual scalability We used in particular text labels as cues, since they are a natural choice for representing textual content. With the semantics of terms depicted by the labels, human users can easily associate and access a wealth of background information, which makes them much more powerful from other visual cues. The interpretability of the focused documents increases with the number of depicted terms, making it even possible to disambiguate complex content and develop an idea of the underlying document set. Even if we use a vector space representation and the shown labels reflect (weighted) dimensions of the document collection under the lens, their effectiveness in conveying information seems to be higher than we experienced with quantitative values of structured information. At the same time we introduce uncertainty by presenting the labels outside their original contexts. This poses the risk of misinterpreting labels and wrong associations. Different techniques of extracting terms might counteract this issue. And by changing lens size and position, additional information can be quickly explored for correcting inaccurate interpretations. The quality and interpretability of characterizations using labels also depends on the length of texts, their number, type, and their spatial distribution. Short texts and content-based spatial distributions of documents 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 as part of our future research. How quickly text labels can be perceived and interpreted by a human analyst as opposed to other visual representations is a different issue. Through the techniques we employ, the labels remain relatively stable if the lens is moved, but this of course depends, on movement speed as well as on the smoothness of the underlying document spatialization. When using DocuCompass ourselves, we had the impression that labels of equal size can be perceived more quickly than ones that encode additional aspects in size, which lead to the decision of representing prominence or importance to the labels sequence. We assume that keeping the (starting) position of the labels stable in relation to the lens supports their quick interpretation. However, further assessment of the shown number and size of labels is required to optimize the visual cues for perception.

In addition to a lens-based approach, brushing techniques such as using a rectangle or polygon brush could be equipped with similar visual cues and analysis techniques. They could offer the benefit of user defined shapes when an adaptation to complex document distributions in the spatialization is useful. Movable lenses have advantages for continuous exploration tasks, we aim for with our approach.

7 Conclusion

With DocuCompass we provide a flexible and easy-to-use exploration method for document spatializations. It fills a gap between visualization and interaction techniques that provide large scale overview and detailed inspection of text corpora. DocuCompass is therefore a powerful complement to many visual analytics approaches which profit from continuous exploration as part of text foraging tasks. The initial user feedback we collected supports this claim. Besides the analysis techniques described in this work, DocuCompass can be easily extended with almost any text analysis procedure. It is thus flexible and adaptable to different text types and analysis tasks. At the moment, we offer a limited set of visual cues for interpreting documents under exploration. While we consider text labels as the most informative means, this set can certainly be expanded in the future. As initial information needs form, exploration gets more targeted. This is supported which with different navigation aids. We did not discuss the transition from our exploration to more details-on-demand. A straightforward way for this can be the depiction of focused documents' details in a separate view. Making this transition as smooth and effective as possible will be part of future research endeavors.

Acknowledgements

We thank Dennis Thom and Harald Bosch who developed early lens versions for their approaches. We also thank the evaluators for their valuable feedback. This work was partly funded by the German Federal Ministry of Education and Research (BMBF) as part of the Center for Reflected Text Analysis 'CRETA' at University of Stuttgart and by the German Research Foundation (DFG) as part of the priority program 1335 'Scalable Visual Analytics'

References

[1] E. Bertini, M. Rigamonti, and D. Lalanne. Extended Excentric Labeling. *Comput. Graph. Forum*, 28(3):927–934, 2009.

[2] E. A. Bier, M. C. Stone, K. Pier, W. Buxton, and T. D. DeRose. Toolglass and Magic Lenses: The See-through Interface. In *Proc. Annu. Conf. on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 73–80, New York, NY, USA, 1993. ACM.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[4] 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. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2022–2031, Dec 2013.

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

[6] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[7] J. Choo, C. Lee, C. K. Reddy, and H. Park. UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Trans. Vis. Comput. Graph.*, 19(12):1992–2001, Dec 2013.

[8] J. Chuang, C. D. Manning, and J. Heer. "Without the Clutter of Unimportant Words": Descriptive keyphrases for text visualization. *ACM Trans. Comput. Hum. Interact.*, 19(3):19:1–19:29, Oct. 2012.

[9] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and Trust: Designing Model-driven Visualizations for Text Analysis. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, CHI '12, pages 443–452, New York, NY, USA, 2012. ACM.

[10] A. Cockburn, A. Karlson, and B. B. Bederson. A Review of Overview+Detail, Zooming, and Focus+Context Interfaces. *ACM Comput. Surv.*, 41(1):2:1–2:31, Jan. 2009.

[11] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. IEEE Conf. on Visual Analytics Science and Technology (VAST)*, pages 91–98, 2009.

[12] M. Correll, M. Witmore, and M. Gleicher. Exploring Collections of Tagged Text for Literary Scholarship. *Comput. Graph. Forum*, 30(3):731–740, 2011.

[13] W. Cui, S. Liu, Z. Wu, and H. Wei. How Hierarchical Topics Evolve in Large Text Corpora. *IEEE Trans. Vis. Comput. Graph.*, 20(12):2281–2290, Dec 2014.

[14] 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.

[15] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2002–2011, Dec 2013.

[16] A. Endert, R. Burtner, N. Cramer, R. Perko, S. Hampton, and K. Cook. Typograph: Multiscale spatial exploration of text documents. In *Proc. IEEE Int. Conf. on Big Data (Big Data)*, pages 17–24, Oct 2013.

[17] A. Endert, P. Fiaux, and C. North. Semantic Interaction for Visual Text Analytics. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, CHI '12, pages 473–482, New York, NY, USA, 2012. ACM.

[18] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. Observation-level interaction with statistical models for visual analytics. In *Proc. IEEE Conf. on Visual Analytics Science and Technology (VAST)*, pages 121–130, Oct 2011.

[19] G. W. Furnas. Generalized Fisheye Views. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, CHI '86, pages 16–23, New York, NY, USA, 1986. ACM.

[20] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Trans. Vis. Comput. Graph.*, 8(1):9–20, Jan 2002.

[21] M. A. Hearst. *Search User Interfaces*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

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

[23] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual Classifier Training for Text Document Retrieval. *IEEE Trans. Vis. Comput. Graph.*, 18(12):2839–2848, Dec 2012.

[24] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word cloud explorer: Text analytics based on word clouds. In *Proc. Hawaii Int. Conf. on System Sciences (HICSS)*, pages 1833–1842, Jan 2014.

[25] 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.

[26] A. Judea, H. Schütze, and S. Brügmann. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proc. COLING*, pages 290–300, 2014.

[27] E. Kandogan. Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In *Proc. IEEE Conf. on Visual Analytics Science and Technology (VAST)*, pages 73–82, Oct 2012.

[28] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[29] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: a new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[30] 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. *J. Exp. Soc. Psychol.*, 49(4):764 – 766, 2013.

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

[32] M. Luboschik, H. Schumann, and H. Cords. Particle-based labeling: Fast point-feature labeling without obscuring other visual features. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1237–1244, Nov 2008.

[33] 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 *Proc. IEEE Conf. on Visual Analytics Science and Technology (VAST)*, pages 181–190, Oct 2011.

[34] J. D. Mackinlay, G. G. Robertson, and S. K. Card. The Perspective Wall: Detail and Context Smoothly Integrated. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, CHI '91, pages 173–176, New York, NY, USA, 1991. ACM.

[35] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[36] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proc. ACL*, pages 55–60, 2014.

[37] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

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

[39] C. North. Toward measuring visualization insight. *IEEE Trans. Vis. Comput. Graph.*, 26(3):6–9, May 2006.

[40] 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. *IEEE Trans. Vis. Comput. Graph.*, 14(3):564–575, May 2008.

[41] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. Int. Conf. on intelligence analysis*, volume 5, pages 2–4, 2005.

[42] C. Plaisant. The Challenge of Information Visualization Evaluation. In *Proc. Working Conf. on Advanced Visual Interfaces*, AVI '04, pages 109–116, New York, NY, USA, 2004. ACM.

[43] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proc. Workshop on Comparing Corpora*, pages 1–6, 2000.

[44] G. G. Robertson and J. D. Mackinlay. The Document Lens. In *Proc. ACM. Symp. User Interface Software and Technology (UIST)*, UIST '93, pages 101–108, New York, NY, USA, 1993. ACM.

[45] A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[46] T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Inf. Vis.*, 9(3):181–193, 2010.

[47] R. Spence and M. Apperley. Data base navigation: an office environment for the professional. *Behav. Inf. Technol.*, 1(1):43–54, 1982.

[48] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions. *IEEE Trans. Vis. Comput. Graph.*, 22(1):629–638, 2016.

[49] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. *Inf. Vis.*, 7(2):118–132, 2008.

[50] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Proc. IEEE Pacific Visualization Symposium (PacificVis)*, pages 41–48. IEEE, 2012.

[51] 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, *Proc. EuroVis - STARs*. The Eurographics Association, 2014.

[52] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9:2579–2605, 2008.

[53] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3 edition, 2012.

[54] 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 *Proc. IEEE Symp. Information Visualization*, pages 51–58, Oct 1995.