MultiCloud: Interactive Word Cloud Visualization for Multiple Texts

Markus John[*]
Institute for Visualization and Interactive Systems (VIS), University of Stuttgart

Eduard Marbach[†]
Institute for Visualization and Interactive Systems (VIS) University of Stuttgart

Steffen Lohmann[‡]
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)

Florian Heimerl [§]
Department of Computer Sciences, University of Wisconsin-Madison

Thomas Ertl [¶]
Institute for Visualization and Interactive Systems (VIS) University of Stuttgart

Abstract

Word Clouds have gained an impressive momentum for summarizing text documents in the last years. They visually communicate in a clear and descriptive way the most frequent words of a text. However, there are only very few word cloud visualizations that support a contrastive analysis of multiple documents. The available approaches provide comparable overviews of the documents, but have shortcomings regarding the layout, readability, and use of white space. To tackle these challenges, we propose MultiCloud, an approach to visualize multiple documents within a single word cloud in a comprehensible and visually appealing way. MultiCloud comprises several parameters and visual representations that enable users to alter the word cloud visualization in different aspects. Users can set parameters to optimize the usage of available space to get a visual representation that provides an easy visual association of words with the different documents. We evaluated MultiCloud with visualization researchers and a group of domain experts comprising five humanities scholars.

Keywords: Visual text analysis, document analysis, word cloud.

Index Terms: Human-centered computing—Visualization—Visualization techniques; Human-centered computing—Visualization—Visualization application domains

1 Introduction

In recent years, word clouds have attracted a lot of attention. Typically, they are used to abstract a text by providing an overview of the most frequently used words. Visual abstractions of text can convey valuable insight and support users in getting a basic understanding of the information a document contains without reading the whole text [17]. Despite their limitations, word clouds are often considered an intuitive visual abstraction [13, 35]. They can provide a valuable starting point for further analysis and have been successfully applied in many different domains, ranging from digital humanities [16] to social media [23, 28] and patent analysis [38].

Word clouds make use of different visual variables. Font size typically encodes relative importance or frequency of the words in a text. Font color is sometimes used to encode additional information, for example, to indicate part-of-speech tags or semantic meanings of words. Word position in the cloud can be used to indicate relationships, for instance, by grouping clusters of related words [25]. Several works deal with the optimization of word cloud layouts in order to compute more effective, informative, or comprehensible representations (see Section 2).

As of yet, however, very few approaches are available that display multiple text documents in a single word cloud [4, 24]. While the available approaches can visualize differences and commonalities between a set of documents, they have shortcomings with regard to the layout, readability, and use of white space: The approaches lack a flexible and comprehensive layout that clearly communicates the composition of the word cloud. The readability is reduced, as the words consume considerable space and can quickly become unreadable if they overlap. In addition, the works do not make optimal use of the available space, resulting in word clouds that are aesthetically less appealing.

In this work, we propose an improved word cloud visualization that depicts a document set in a merged view. It provides different parameters and visual representations to influence the layout of the cloud. The options allow users to customize the word cloud and overcome the main limitations of previous work. For example, they can modify the spatial arrangement of the words to make more efficient use of the available space. Furthermore, interaction techniques enable the users to further analyze the word cloud visualization and get details on demand.

MultiCloud has been created in the context of the digital humanities project CRETA [7]. The project goal is the development of technical approaches and a general work flow methodology for text analysis within the Digital Humanities. Our humanities collaborators are interested in the analysis of novels consisting of several books. Based on discussions with them, we have derived practical analysis scenarios and tasks. For example, it is important for them to get an overview of prominent topics, differences and commonalities of a set of documents. To support such comparative analyses, we decided to develop a word cloud view that makes differences and commonalities in word use immediately visible, since word clouds – despite all known drawbacks – are easy to understand and the humanities scholars are familiar with them. The presented approach should provide first feedback and serves as a basis for further development of our approach in close cooperation with the humanities scholars.

With MultiCloud, we extend previous work into several directions and provide an approach that visualizes a set of text documents in a single merged word cloud. It resolves shortcomings regarding layout, readability, and white space.

The rest of the paper is structured as follows: Section 2 summarizes related work, before our approach is presented in Section 3. This is followed by a usage scenario demonstrating the applicability and usefulness of MultiCloud in Section 3.3.2. In Section 5, MultiCloud is evaluated with visualization experts. In addition, we report on user feedback that has arisen in a focus group workshop involving five humanities scholars. Section 6 provides a discussion and outlines future work, before we conclude with Section 7.

[*] e-mail: markus.john@vis.uni-stuttgart.de
[†] e-mail: eduard.marbach@vis.uni-stuttgart.de
[‡] e-mail: steffen.lohmann@iais.fraunhofer.de
[§] e-mail: heimerl@cs.wisc.edu
[¶] e-mail: thomas.ertl@vis.uni-stuttgart.de

Graphics Interface Conference 2018
8-11 May, Toronto, Ontario, Canada

2 Related Work

Since our approach is an extension of the basic word cloud visualization technique, we first summarize works that investigate the effectiveness and visual perception of word clouds. Then, we look into related approaches on performing text analyses based on word clouds. Lastly, we review previous work that visualize multiple documents in a merged word cloud and highlight the extensions and improvements of our approach compared to those works.

2.1 Effectiveness and Perception of Word Clouds

In the last decade, there have been several attempts to examine the effectiveness and perception of word clouds [1]. Bateman et al. [2] present the results of a user study which indicate that the font size, weight, and color in word clouds have the largest effect on the users' attention. They have also determined that large words in the center of the cloud receive most user attention, as it was confirmed by Lohmann et al. [25] using eyetracking technology.

Several research works compared word clouds with unweighted lists and other user interface elements [12,25,27]. When searching for a specific word, alphabetically ordered word clouds have been found to be less effective than alphabetically ordered lists without any weighting (i.e., with words in a uniform font size). However, frequently occurring words can be spotted more quickly in word clouds due their larger font size.

Sinclair and Cardew-Hall [29] compared word clouds with search interfaces. They found that most users prefer a search box to find specific words, but like word clouds for more open-ended tasks, as they can provide a quick overview of a text document and serve as a starting point for further analyses.

Felix et al. [10] conducted a number of user studies that aimed at exploring the visual design space of word clouds. The studies focused on the spatial layout and value encoding and include several combinations of them. Based on the results, they defined guidelines on how to design effective word clouds, and emphasize that the performance of word clouds highly depends on the task they are used for.

2.2 Improvements and Extensions of Word Clouds

Apart from research work, freely available web services, such as Wordle [37], Tagul [31], or Tagxedo [32], have emerged that enable the creation of appealing word clouds. Users can select different options to customize the visualizations. For example, they can change the word orientation, color, or bounding box (i.e., general shape) of the word clouds. Further layout algorithms and customization options have been proposed in ManiWordle [21] and Rolled-out Wordles [30]. However, these design-oriented approaches do not include features to compare and analyze different text documents.

Lee et al. [22] presented SparkClouds, which displays sparklines below the words to represent changes in word use over time. A related approach was introduced by Lohmann et al. [23], who use histograms with visual highlighting of co-occurrences to indicate time-dependent word relations. Gambetta and Véronis [11] proposed Tree Clouds that combine word clouds with trees to depict semantic relationships extracted from the text. Prefix Tag Clouds [5] uses a prefix tree to visually group different word forms by color and space. This allows users to easily identify and compare the used word forms in the cloud. Other approaches focus on improving the layout and use of white space in word clouds. For instance, Kaser and Lemire [19] apply different techniques, such as slicing trees and nested tables, to optimize the distribution of white space in word clouds displayed on the web.

Word clouds do not have to stop at the stage of static visualizations, which they are still largely used for. Adding interaction can be advantageous in several directions. For example, there are interactive word cloud visualizations that enable users to select multiple words to highlight relations in the cloud or related elements in other views, respectively [13,20,38]. In addition, a couple of approaches allow users to filter different word forms, such as verbs or nouns, or selected words [9,13,35]. EdWordle [36] enables users to move and edit words in the cloud while preserving the neighborhoods of other words. Thus, users can update the word cloud and create compact layouts based on their needs.

Finally, there are approaches that offer a possibility to inspect the corresponding text passages of selected words through interaction [13,18,35]. This way, users get an overview of the text content with the aid of the word cloud, and can further analyze the detailed context in the text view.

2.3 Word Clouds for Multiple Text Documents

There are only a handful approaches that use word cloud visualizations in order to differentiate among facets within a text corpus. Parallel Tag Clouds [6] combine the ideas of word clouds, small multiples, and parallel coordinates to provide overviews of a document collection. However, they represent the same words multiple times and the different texts are only implicitly indicated by font color and style. Thus, it is difficult to compare and understand the commonalities and differences of the texts.

Viegas et al. [34] and Diakopoulos et al. [8] depict words from different texts in a merged word cloud. They indicate the text source of each word simply by its font color. Both approaches have the limitation that they mainly scale for the comparison of two or three documents, as the same words are depicted multiple times (once for each document they occur in).

Jänicke et al. [16] presented TagSpheres which visualizes various types of text hierarchies in a compact word cloud. TagSpheres uses color and stacked bar charts along with each word in order to indicate its relevance in different categories. However, TagSpheres has been designed primarily to visualize textual summaries that comprise hierarchical information, whereas we are interested in comparing commonalities and differences between text documents.

Most closely related to our approach are RadCloud [4] and ConcentriCloud [24]. RadCloud also shows extracted words from a set of text documents in a single word cloud. It provides a circular shape and the different documents are referenced on the border of the layout. RadCloud uses a force-directed layout to place words as close as possible to the computed positions. It makes use of color and stacked bar charts to represent how relevant a word is in each document. However, the approach has drawbacks regarding an appealing overview and efficient use of screen space, since the algorithm tends to generate a lot of white space and many small words, which are difficult to read.

ConcentriCloud addresses these shortcomings and introduces a space-filling approach. The different documents are represented on the outermost circle and the merged ones on inner circles (cf. Figure 7). However, the approach has a couple of limitations: For instance, it is not possible to analyze differences and commonalities of documents arranged at opposite positions of the outer circle. In addition, if words cannot be placed within the available space, they are omitted from the word cloud and ConcentriCloud tries to place the word with the next highest frequency value. Thus, it can happen that words which are very frequent and central in the text documents are skipped and not shown at all in the word cloud. Furthermore, both approaches have a fixed bounding box of circular shape and provide only few options to customize and optimize the word cloud.

3 MultiCloud

To tackle the shortcomings of previous works, we defined four design goals: (**G**1) facilitating users in understanding the commonalities and differences of multiple text documents; (**G**2) distributing the words in a way that makes efficient use of the available space and reaches an appealing overview of the content; (**G**3) aiming to represent the most relevant words in the visualization without skipping any

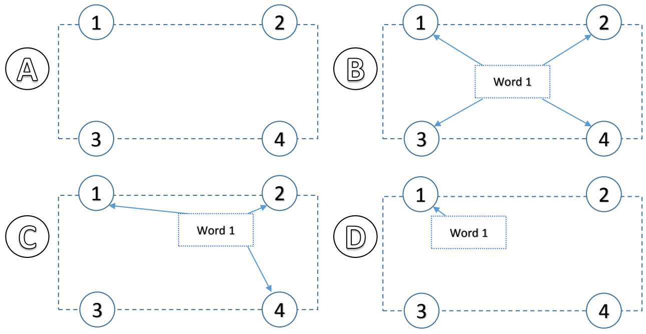

Figure 1: Schematic representation of the layout: four documents are uniformly distributed as fixed points on the border of the layout: (A) without any weighting, (B) with an equal weighting of a word to all documents, (C) without a weighting to document 3 and a strong weighting to document 2, (D) with only a weighting to document 1.

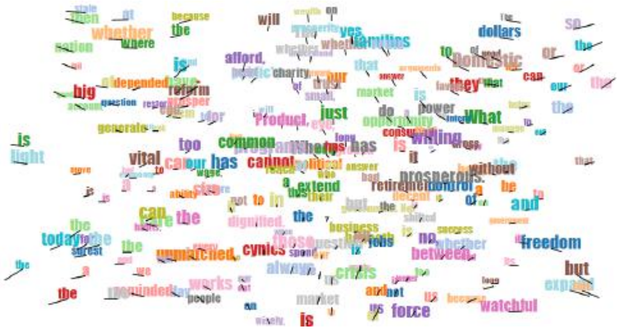

Figure 2: First attempt to display the words using a force-based layout. The black lines indicate the deviations of the actual word positions to the calculated ones caused by the forces.

candidates; (**G**4) offering several possibilities to customize the word cloud visualization, such as different layout shapes or options to modify the placement of words.

MultiCloud uses a force-based layout in combination with a collision detection algorithm to tackle the shortcomings of earlier works and to meet the above listed design goals. In the following, we briefly report on the text processing steps, before we detail the visual design and implementation of MultiCloud, and describe the parameters implemented to modify the word cloud visualization.

3.1 Text Processing

Once a number of documents is loaded into our MultiCloud implementation (as plain text or in EPUB format), the documents are processed in a number of linguistic analysis steps. We use the Apache Lucene Core library[1] to tokenize the text and apply a weighting scheme. The weightings are used to indicate the relevance of the visualized words for the different documents. We use the popular *term frequency (tf)* and *term frequency–inverse document frequency (tf–idf)* measures to compute the weightings of the words. For *tf-idf*, we count the word occurrences in the individual documents *tf*, and counterbalance this value with the words' occurrences across all documents, using the logarithm of the words' *inverse document frequency* (idf). Thus, less emphasis is put on words frequently used throughout the whole document set, which results in frequent words that have minor discriminating information being effectively removed from the word cloud visualization. In addition, we provide a stop word list that can optionally be activated to filter even more words that usually do not carry relevant information.

Furthermore, we use the Stanford CoreNLP framework[2] to extract more meaningful information. In particular, we run part-of-speech (POS) tagging to classify words, for example, as verbs, adjectives, or nouns, and apply named entity recognition (NER) to extract entities such as people and places from the text documents. Based on the NLP results, we enable the filtering of the word cloud by the detected POS tags or named entities.

3.2 Visual Design and Implementation

We implemented MultiCloud as a web-based visualization that uses standard web technologies and runs with any modern web browser supporting HTML5, SVG, CSS, and JavaScript. We applied a force-based layout, using D3 [3], to spatially arrange the extracted words in the cloud. The layout consists of nodes and edges, where nodes represent the words and the edges encode the relevance weightings. The different documents are depicted as fixed points on the border of the word cloud, as illustrated with the numbers in Figure 1ⓐ.

As a next step, we distribute the words on the canvas using the fixed document points as the coordination system. We sort the words per document by their weighting and define an edge for each word to the respective documents where it appears. That is, for each weighting value > 0, we create an edge to the document.

An example is shown in Figure 1ⓑ: the blue circles represent the fixed document points, the dashed strokes the general word cloud shape, and the arrows the edges. If the weight distribution of a word is equal in all documents (i.e., if the word appears exactly the same number of times in all documents), the word is placed in the center of the canvas, as illustrated in Figure 1ⓑ. If the weights are not equally distributed (which is usually the case), the word is moved in the direction of the documents where it has the strongest weightings by the force-based layout. For example, if a word does not occur in a particular document (weight = 0), no edge is added to that document point, as depicted in Figure 1ⓒ. In case a word occurs only in one of the documents, the word is placed close to the respective document point, as depicted in Figure 1ⓓ.

If the number of words do not fill the existing area sufficiently, we alter the forces of the layout to distribute the words optimally. This is aesthetically more pleasing than a word cloud with a lot of (and unbalanced) white space. However, despite the changes, the visual mapping must be maintained. Therefore, we use a font scaling to fill the existing space more efficiently.

To obtain an overlap-free and well-distributed layout, we tried to find an appropriate choice of the parameters, such as the font size scaling and maximum space utilization. Despite different variations of such parameters, there are still many overlaps, as depicted in Figure 2. However, it is recognizable that the positions of the words differ only slightly from the original positions, as indicated through the black lines. To avoid overlapping words, we developed an approach that is detailed in the following.

At first, we implemented the approach using the JavaScript library *cola.js*[3], which provides constraint-based optimization techniques for force-based graphs and diagrams. It can be integrated into D3 and allows users to specify constraints such as alignments or groupings of nodes. The approach automatically generates constraints to avoid overlapping nodes. However, the generated layout works only well for small word graphs and cannot be arranged in a particular shape, as depicted in Figure 3ⓐ. Through awkward positions whole blocks may move, as depicted in Figure 3ⓑ. We tried to modify the constraints in order to solve this problem. However, it did not lead to a satisfactory result, because unfortunate distributions can cause inappropriate layouts.

[1] https://lucene.apache.org/core/

[2] http://nlp.stanford.edu/software/corenlp.shtml

[3] http://marvl.infotech.monash.edu/webcola/

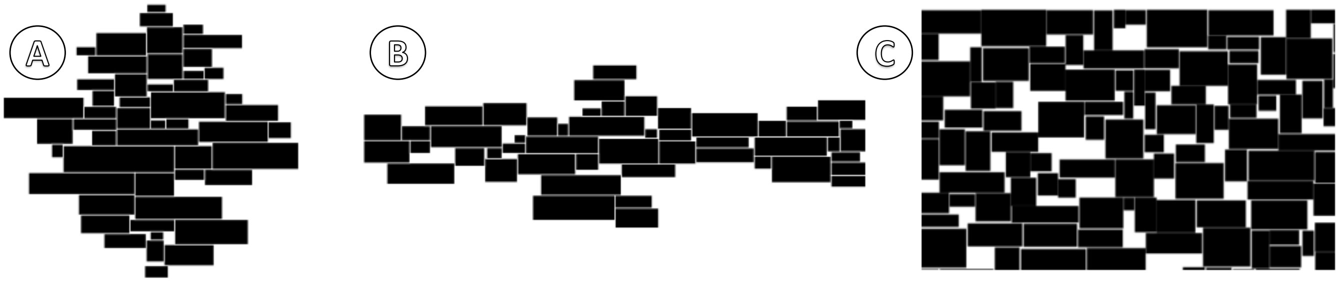

Figure 3: (A) A prototypical implementation based on *cola.js* shows promising results for small graphs of words. (B) However, through awkward parameters and positions, the layout arranges itself into an elongated form instead of a compact layout. (C) After discarding the *cola.js* library, we implemented another approach that generates a more space and shape filling layout.

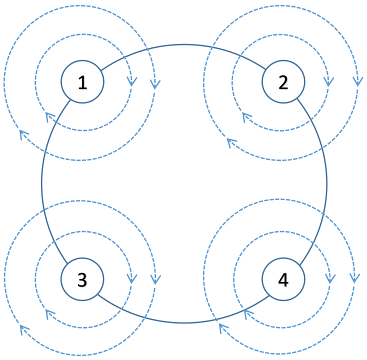

Figure 4: Illustration of our positioning algorithm: The inner circle indicates the layout shape, while the other circles represent the sampling points for the analyzed documents (four documents in this example).

Therefore, we discarded the *cola.js* library and developed the idea to place the words (sorted by their weightings) as close as possible to the document points. If there is an overlap of words, the algorithm scans the nearest area and tries to find a free position, as depicted in Figure 4. This way, we could reach promising and space-filling results, as depicted in Figure 3ⓒ.

Eventually, this approach led to the final design, which works as follows: As a first step, we calculate the weightings of the words by using either *tf* or *tf-idf*. In addition, we apply a force-based layout, using D3 [3], to determine the initial positions of the words. Subsequently, we place one word after another for each document. For example, we take and place the highest ranked word of document one, then place the most relevant word for document two, etc. That way, strongly associated words are placed close to the document points. The algorithm runs until no words are any longer available in the lists and tries to find a free position for each word. However, there may be cases in which words have no place in their assigned area. This results in words being drawn at different positions than intended. To prevent such effects, we integrated a tolerance range in which a placement of a word is still acceptable. Additionally, we implemented an expert mode that shows the deviation errors of the word positions in the cloud, i.e., the deviations of the actual word positions to the originally computed ones. For each word, a line is drawn indicating the deviation error, as it is shown in Figure 5. The length of the line represents the error strength: the longer the line, the larger the deviation. Using this mode, expert users can determine deviation errors easily and have a chance to fix them by adjusting the settings in case the errors get too large.

In general, we enable all users to modify the visual layout. In the following, we will detail these possibilities.

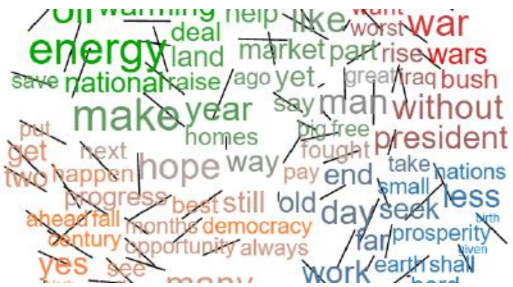

Figure 5: The deviation is indicated by a line from the actual to the computed word position. The longer the line, the larger the deviation.

3.3 Parameters and Visual Representations

We provide several parameters and options to influence the layout of the word cloud, such as different layout shapes, options to modify the placement of words, or to alter the intersections of the different documents. In addition, we offer interaction possibilities and visual representations that help to better understand and explore the word cloud visualization.

3.3.1 Layout Shapes and Placement of Words

Our approach supports three main layout shapes: **rectangle**, **circle**, and **ellipsoid**. In addition, we provide three options to change the placement of the words and thus the visual layout: Users can choose between the options to 1) use the maximum space, 2) place all words, and 3) center the words:

1. Maximum space utilization: This option allows users to modify the word cloud in a way that the available space is used as optimally as possible. To reach this goal, we iteratively increase the font sizes of the words until the threshold where the required space gets greater than the available space. This way, the layout generates as little whitespace as possible.

2. Placement of all words: This option enables users to enforce that all words are represented and no word is omitted. Depending on the number of words and the tolerance range, a situation may occur where not all words can be placed. Accordingly, we tailored the algorithm so that the font sizes of the words are iteratively decreased as long as any word is omitted.

3. Centralization of words: With this option, it is possible to reduce the white space in the center of the word cloud. It sorts the words starting from the center and then iteratively moves the words in the calculated directions until a collision occurs. This method is similar to the prototypical implementation shown in Figure 3ⓒ. One general drawback of this method is that the actual positions of the words can deviate strongly from the originally calculated positions.

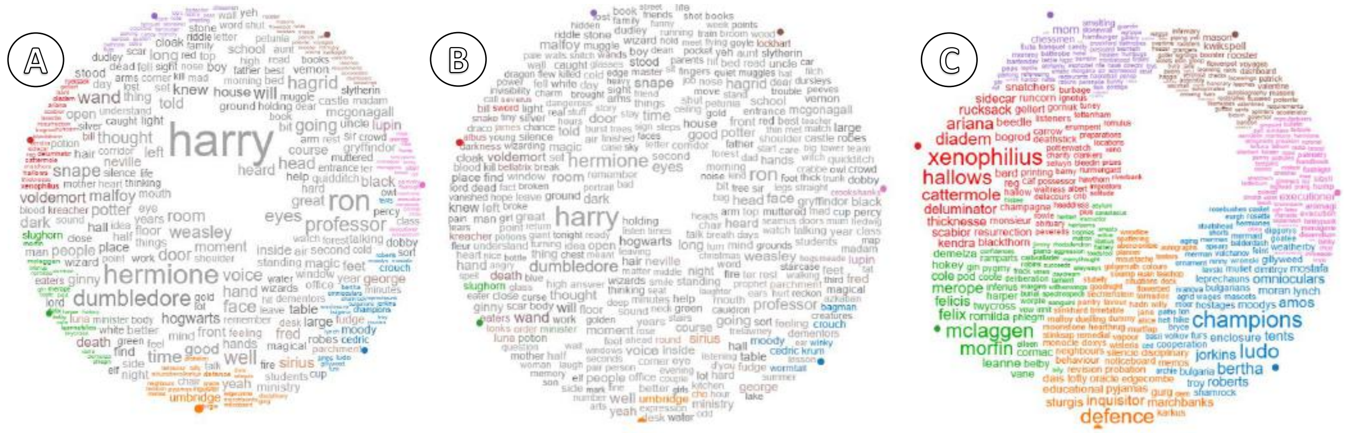

Figure 6: Three different word cloud variations using (A) a mixed word relevance, (B) only words that occur in multiple documents, and (C) words, which only appear in the individual documents.

3.3.2 Font Size and Scaling

The initial font sizes of the words are scaled with their occurrence frequency, depending on the overall frequency distribution in the documents by using a square root function. Relevant words are scaled up to a large font and stand out, while irrelevant words are displayed smaller. This way, users get a first impression of the contents of the document set. In addition, we enable users to alter the minimum and maximum font size used in the cloud.

Next to the square root scaling, we offer three alternative functions for scaling the font sizes: linear, logarithmic, and quadratic. The linear and quadratic scaling lead to stronger differences in the font sizes of the largest and smallest words. In contrast, the logarithmic and square root scaling reduce the visual lie factor [33] and provide a more even distribution of the font sizes.

3.3.3 Word Relevance

In addition to the position in the layout, we use color to indicate how relevant each word is in the different text documents. To distinguish the individual documents and word associations, discrete colors are required. We use predefined color schemes of ColorBrewer[4] for a qualitative, color-blind safe, and print-friendly color assignment. We assign a unique color to each of the loaded documents. Subsequently, we assign each word the color of the document where the word most often occurs. Furthermore, we use color saturation to indicate how strong the word is associated with the document, i.e., lower saturation means lower association.

Moreover, users have the possibility to control the number of displayed words: They can set both the number of displayed words that occur in more than one document and the number of displayed words that occur only in an individual document. Both options can be combined and trigger an immediate update of the visualization.

3.3.4 Interaction

To allow for a better exploration of the word cloud, we added several interaction techniques. We support common user interactions, such as panning, zooming, or rearranging. When hovering over a word in the cloud, a tooltip shows a pie chart with the word frequency distribution in the different documents. By clicking on a word, additional information is shown in a separate view, such as the POS tag or the individual and overall occurrences of the word. Further, users can click on a document to highlight all words that have an assignment to the document higher than a certain threshold (in the default case: 0.5). That way, users can easily get an overview of the most relevant words of a document. In addition to the aforementioned options to modify the word cloud, users can filter the words by the different POS tags, as it was mentioned before.

4 Usage Scenario

In the following, we present a usage scenario that demonstrates the applicability and usefulness of our approach. In the scenario, we analyze the novel series "Harry Potter" written by J.K. Rowling. The storyline is about the life and adventures of the young wizard Harry Potter, the protagonist, and his struggles against the dark wizard Lord Voldemort. To explore the commonalities and differences of the seven volumes of the series, we defined the following goals: (1) We aim to investigate a combined word cloud that represents both the words of the individual novels and of all seven volumes to get a first idea of their distribution and the most prominent words. (2) Next, we want to identify words that represent the document collection of Harry Potter. This way, we get an overview of words that occur in several of the novels. (3) Last, we want to inspect words that represent the individual volumes. Thus, entities or events that are mentioned in the individual volumes can be recognized.

As a first step, we load the volumes of "Harry Potter" into our MultiCloud implementation. After the novels have been processed, we choose a circular layout and the options *maximum space utilization* and *placement of all words* to get a clear and comprehensive word cloud, as depicted in Figure 6ⓐ. In addition, we set the number of words that should be extracted from all documents to 200 words, and those which should be extracted from the individual volumes to 20 words. The words colored in gray, which occur in all seven volumes, are arranged in the center of the word cloud.

The white space in the center of the cloud results from the tolerance range (see Section 3.2), which prevents that the words are placed at arbitrary positions. While analyzing the word cloud, we can easily identify the main characters, such as Harry, Hermine, Ron, or Dumbledore. In addition, we can determine characters that occur only in one of the volumes, such as Xenophilius Lovegood (red document) or Bartemius Crouch (blue document). By hovering over the characters, a tooltip shows additional information, such as the exact occurrences in the different volumes. This way, for example, we find out that Xenophilius Lovegood only occur in "Harry Potter and the Deathly Hallows – Part 1" (red document).

To investigate words that represent the whole document collection, we change the word relevance in the options menu. We set the number of words that only occur in one volume to zero and increase the number of words that are mentioned in multiple volumes. The resulting word cloud is shown in Figure 6ⓑ. Again, we can easily

[4] http://colorbrewer2.org

recognize frequently occurring entities or events in the word cloud. For example, we can identify characters that are mentioned in several documents, but mainly occur in one volume, such as Remus Lupin (purple document), Doloras Umbridge (orange document), or Horace Slughorn (green document).

In the next step, we change the word relevance again, since we are interested in examining those words that represent individual novels. For this purpose, we set the number of words that occur only in individual novels to 70 words, and the number of words that occur in more than one volume to zero. By exploring the word cloud (Figure 6ⓒ), we get a quick overview of the words prominent in the different volumes. The document words are distributed at the border of the circle up to the center. Since the most frequent words of the whole series are not included, the document words come to light and are scaled higher. Thus, we can discover several specific entities, events, and objects of the individual volumes, such as Griselda Marchbanks, head of the Wizarding Examinations Authority (orange document), the Triwizard Champions Tournament (blue document), or the Ravenclaw's Diadem object, which is sought by Harry Potter (red document). In addition, we can confirm that Xenophilius Lovegood plays an important role in the penultimate novel of the Harry Potter series (red document).

The usage scenario shows that MultiCloud can facilitate users in exploring multiple documents in a compact and contrastive way. We offer several possibilities to change the layout of the word clouds in order to investigate different questions. MultiCloud can therefore support users in gaining insights as well as in generating and confirming hypotheses.

5 Expert Evaluation

To evaluate MultiCloud, we conducted a qualitative study with visualization experts, based on a comparative analysis with the previous work ConcentriCloud (cf. Section 2). In addition, we presented our approach in a focus group workshop, which involved five humanities scholars.

5.1 Evaluation with Visualization Experts

As a basis for the expert study, we loaded the seven volumes of the Harry Potter series in the implementations of MultiCloud and ConcentriCloud, and prepared a number of evaluation tasks. For example, participants had to find one of the most relevant words in volume five, a word that only occurs in the fourth volume, how often the name "Harry" appears in the first volume, or they had to solve a search task to find a specific word in the cloud. Furthermore, we designed a final questionnaire with open and Likert scale questions, and the participants had the chance to give final remarks.

ConcentriCloud implementation: ConcentriCloud is also based on a word cloud visualization and uses a concentric layout to show differences and commonalities between multiple documents (cf. Section 2). As the only layout, it offers an ellipsoid word cloud shape, where the individual documents are represented on the outermost circle and the merged ones on inner circles, as depicted in Figure 7. The approach attempts to place the words within the bounding box of the respective word clouds. Initially, the most frequent words are placed and then the algorithm continues with words of decreasing frequencies. This way, the most frequent words are placed in the center of the respective word cloud, such as "Harry". However, if words cannot be placed within the respective bounding box, they are omitted from the word cloud and the algorithm tries to place the next frequently occurring word into the word cloud. A separate word list, which contains all extracted words with their occurrences complements the word cloud view. In addition, interactive features similar to ours provide details on demand. For example, users can filter the word cloud by POS tags to show only nouns. If a users hovers over a word, the respective documents are highlighted and a tooltip shows the overall and individual numbers of its occurrences in the document set.

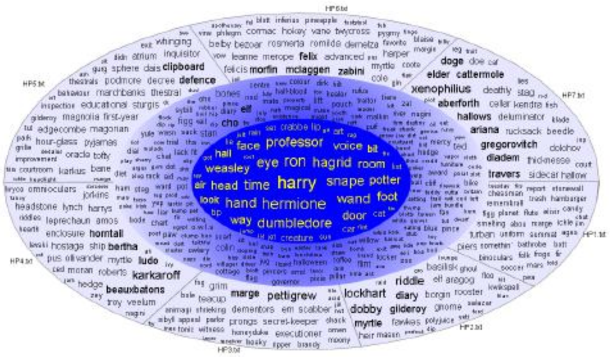

Figure 7: ConcentriCloud visualization for the Harry Potter series.

Participants and Procedure: We recruited eight visualization experts, one female and seven males. Their average age was 30 years (min 27, max 32). All participants had a strong background in information visualization, and most of them were familiar with word clouds. We used a 10-point (0 = no knowledge and 10 = expert) Likert scale to determine the previous knowledge the participants had with the visualization of multiple documents (the resulting mean value was 7.25). The individual study took about 45 min., depending on the speed of the participant and the length of the subsequent discussion. We conducted the user study using the thinking-aloud method: Participants were asked to voice their thoughts during the session.

We prepared two document collections for the user study. The first one was used for an introductory session and contained the three books of the "The Lord of the Rings" by J. R. R. Tolkien. The second document collection was the above introduced fantasy series "Harry Potter" that was used for the main study. In addition, we prepared eight tasks, which could be solved with both approaches (see above).

The procedure for each study session was as follows: (1) The participants had to fill out a questionnaire with information about their person, scientific background, and experience with word clouds and the visualization of document collections. (2) Then, we gave each participant a brief introduction into the two visualization approaches on the example of the "Lord of the Rings" trilogy. (3) Afterwards, the participants had to solve four questions (with increasing difficulty) based on the Harry Potter series with each approach. The presentation order of the approaches was counter-balanced between the participants, i.e., four of the experts were shown MultiCloud first, the other four were first exposed to ConcentriCloud. (4) After the last task, the participants had to give feedback about each of the visualizations using Likert scale questionnaires and open questions.

Results: All participants could successfully solve the tasks with both approaches. Seven of the eight visualization experts perceived our approach as aesthetic and more intuitive than ConcentriCloud. One participant considered neither of the two types of word cloud visualization to have the edge over the other, since both were not intuitive from her perspective. At the same time, however, the participant mentioned that this could be due to her lack of experience with text visualization and word clouds.

Five participants thought that the space utilization of our approach is more efficient and clearer than that of ConcentriCloud. This can be attributed to the fact that our approach does not include so many sections and thus the words are better distributed. In addition, the possibility to highlight associated words by clicking on the different documents helps users to recognize and explore document intersections faster.

Four participants noted that the positions of the words in the different intersections have no meaning in ConcentriCloud and that it is thereby difficult to determine how strong the assignment is to the respective documents. They mentioned that our approach solves this problem much better, through the position, the font size, and the color saturation of the words. Thus, it is directly recognizable how often a word occurs and how strongly it belongs to the different documents. Two of the participants especially liked the chosen colors and the color saturation. However, another participant noted that the color coding does not work with many documents. That is true, even though we support the largest color schemes of ColorBrewer.

Three of the eight participants praised the configurability of our approach. They mentioned that it was great to test different options and visual representations and that it helped to analyze the content of the word cloud. In particular, many participants liked the *placement of all words* option, since they were disappointed that it can happen that words could be omitted from the word cloud in ConcentriCloud.

Two further participants remarked that with our approach it is possible to analyze differences and commonalities of opposite documents, whereas ConcentriCloud does not support this. However, our approach is also limited when words occur in opposite documents with the same number of occurrences, since they would be placed in the center of the word cloud. Thus, the impression can arise that words are mentioned in more than the two documents. This was also noticed by two participants which mentioned that ConcentriCloud supports a clearer assignment. Therefore, we implemented interactive features (see Section 3.3.4) to better investigate such cases.

All participants rated the option to show the deviation errors as very helpful to get an impression of the accuracy of the word placement. Some of them mentioned that is especially useful to find optimal settings for the word cloud layout.

The participants were in disagreement regarding the pie charts in the tooltip. Some found them appealing and helpful to get a rough idea of the distribution of the words. Others suggested to use another visualization which supports the comparison of the values better, such as bar charts. Therefore, we decided to provide additional bar charts and let users switch between the two options.

Four participants found that ConcentriCloud is more suitable for tasks where frequencies should be compared, since the approach offers a word list showing the different occurrences of the words. However, our approach could be easily complemented with such a word list.

Several participants suggested specific applications where MultiCloud could be useful, such as to analyze multi-party conversations or the twitter behavior of different users.

Overall, our approach proved to be an improvement compared to ConcentriCloud. The participants liked to work with our approach and the several options to alter the word cloud visualization facilitated their analysis.

5.2 Evaluation with Humanities Focus Group

To evaluate how our approach is accepted by a potential target group, we presented MultiCloud in a focus group workshop that involved five humanities scholars. First, we gave a talk of about 30 minutes to introduce the word cloud layout, the parameter possibilities, and the interactive functions. Since all of the participants were familiar with word clouds, they quickly understood our approach, but had some questions concerning the different parameters. However, the open questions could be quickly answered by practical examples.

The first impression of the humanities scholars was very positive. Three of them mentioned that the visualization provides a good overview to discover differences and commonalities of the document set. One participants praised the many customization possibilities. At the same time, however, two other humanities scholars mentioned that it would be necessary to see the different generated word clouds next to each other for a better comparison. Otherwise, the layout changes and actual influences of the parameters are hard to track. All humanities scholars agreed that the visualization would support their analysis and that it can serve as a starting point for deeper analysis. However, they emphasized that it would still be necessary to work with the text directly in order to compare and inspect text passages in detail.

Overall, the focus group workshop showed that the humanities scholars have great interest in such an approach. For the future, we want to further develop the approach in close cooperation with them. That way, we can respond to their needs and implement specific features and visualizations in a formative process in order to better support their analyses.

6 Discussion and Future Work

This section discusses issues that arose from the expert evaluation, scalability aspects, and further challenges that we would like to address in the future.

A key challenge is scalability when visualizing multiple documents in a merged word cloud. Our approach uses color to distinguish the assignments of the words to the different documents. We use the largest color schemes of ColorBrewer for a qualitative, colorblind safe, and print-friendly color assignment. However, the number of distinct colors is finite, and colors can not be differentiated well at some point. Instead of using color, we could use glyphs, however, they are hard to learn and require additional space, and are thus not a valid alternative.

Another general challenge is the available screen space, since words can quickly become unreadable if they overlap, or words even need to be omitted in the end. One possibility could be to integrate interaction techniques that only display certain words on demand. Furthermore, we could implement focus+context techniques that present detailed information in context or an overview+detail approach, which offers multiple views with different levels of abstractions, for instance, by using some hierachical clustering approach [14].

In the future, we plan to integrate a lemmatization method [26] in the text processing step to reduce each token to a lemma. Lemmatization, for example, can handle irregular forms, such as merging *chose* and *choose*, and brings the word in its base or dictionary form. Additionally, we plan to integrate a coreference resolution technique to find all expressions that refer to the same entity in the document collection. With coreference resolution, we can unify entities automatically, for example, *Harry* and *Mr. Potter* refer to the same entity and provide more information, such as additional names of the characters. An extension related to that would be the detection of multiword expressions, as it can be found in the WordCloudExplorer [13].

During the evaluation, several participants mentioned that it would be very useful to offer a possibility to inspect text passages in detail. Thus, we would like to facilitate a distant and close reading approach [15]. The visual abstraction of the word cloud view provides a starting point for new ideas and hypotheses, while the text view supports a deeper analysis of the content of different documents. Another remark was that users should be able to change the ordering of the documents. This is certainly true and we want to implement such a possibility in the future, since there are some cases where the layout does not provide a clear overview, for example, when a word only occurs in opposite documents. Furthermore, there was a comment that the length of a document could be represented using a chart. We agreed with that and mapped the length of the document to the size of the respective document circle.

In addition, by clicking on the document, our approach shows metadata to the users, such as the title of the book or the author name(s). Another participant suggested to rescale the words after selecting a document, since it is not directly clear how often a word occurs in the selected document.

7 Conclusion

In this work, we presented a flexible approach to show different documents in a merged word cloud visualization. We wanted to overcome limitations of previous work. Therefore, we tackled the challenges to develop an approach that facilitates users in understanding the commonalities and differences of multiple documents, to reduce white space and reach an appealing overview of the content, to represent all important words of the documents, and to offer several possibilities to adapt the word cloud visualization to the users' needs. The usage scenario shows that our approach is effective for the exploration of differences and commonalities between multiple documents. The expert evaluation we performed confirms the need for such an approach and the improvements we reached.

Acknowledgments

This work was funded by the German Federal Ministry of Education and Research (BMBF) as of the Center for Reflected Text Analysis (CRETA) [7] at University of Stuttgart.

References

[1] E. C. Alexander, C.-C. Chang, M. Shimabukuro, S. Franconeri, C. Collins, and M. Gleicher. Perceptual biases in font size as a data encoding. *IEEE Trans. Vis. Comput. Graphics*, 2017.

[2] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: The effect of visual features on tag cloud selections. In *19th ACM Conf. on Hypertext and Hypermedia*, HT '08, pp. 193–202. ACM, 2008.

[3] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011.

[4] M. Burch, S. Lohmann, F. Beck, N. Rodriguez, L. D. Silvestro, and D. Weiskopf. RadCloud: Visualizing multiple texts with merged word clouds. In *18th Int. Conf. on Information Visualisation*, IV '13, pp. 108–113. IEEE, 2014.

[5] M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf. Prefix tag clouds. In *Proc. 17th Int. Conf. Information Visualisation*, IV '13, pp. 45–50. IEEE, 2013.

[6] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *IEEE Symp. on Visual Analytics Science and Technology*, VAST '09, pp. 91–98. IEEE, 2009.

[7] CRETA. Center for reflected text analytics. `https://www.creta.uni-stuttgart.de/en/`.

[8] N. Diakopoulos, D. Elgesem, A. Salway, A. Zhang, and K. Hofl. Compare clouds: Visualizing text corpora to compare media frames. In *IUI Workshop on Visual Text Analytics*, 2015.

[9] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. VisGets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1205–1212, 2008.

[10] C. Felix, S. Franconeri, and E. Bertini. Taking word clouds apart: An empirical investigation of the design space for keyword summaries. *IEEE Trans. Vis. Comput. Graph.*, 24(1):657–666, 2018.

[11] P. Gambette and J. Véronis. Visualising a text with a tree cloud. In *11th IFCS Biennial Conf. and 33rd Annual Conf. Gesell. für Klassifikation e.V.*, pp. 561–569. Springer, 2010.

[12] M. J. Halvey and M. T. Keane. An assessment of tag presentation techniques. In *16th Int. Conf. on World Wide Web*, WWW '07, pp. 1313–1314. ACM, 2007.

[13] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl. Word cloud explorer: Text analytics based on word clouds. In *47th Hawaii Int. Conf. on System Sciences*, HICSS '14, pp. 1833–1842. IEEE, 2014.

[14] D. Herr, Q. Han, S. Lohmann, and T. Ertl. Visual clutter reduction through hierarchy-based projection of high-dimensional labeled data. In *42nd Graphics Interface Conf.*, pp. 109–116. ACM, 2016.

[15] S. Jänicke, G. Franzini, M. F. Cheema, and G. Scheuermann. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In *Eurographics Conf. on Visualization – STARs*, EuroVis '15. Eurographics Association, 2015.

[16] S. Jänicke and G. Scheuermann. On the visualization of hierarchical relations and tree structures with TagSpheres. In *Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics*, VISIGRAPP '16, pp. 199–219. Springer, 2016.

[17] M. John, S. Koch, F. Heimerl, A. Müller, T. Ertl, and J. Kuhn. Interactive visual analysis of german poetics. In *Int. Conf. on Digital Humanities (DH)*, p. 7, 2015.

[18] M. John, S. Lohmann, S. Koch, M. Wörner, and T. Ertl. Visual analysis of character and plot information extracted from narrative text. In *Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics*, VISIGRAPP '16, pp. 220–241. Springer, 2016.

[19] O. Kaser and D. Lemire. Tag-cloud drawing: Algorithms for cloud visualization. In *WWW' 07 Workshop on Tagging and Metadata for Social Information Organization*, 2007.

[20] S. Koch, H. Bosch, M. Giereth, and T. Ertl. Iterative integration of visual insights during scalable patent search and analysis. *IEEE Trans. Vis. Comput. Graphics*, 17(5):557–569, 2011.

[21] K. Koh, B. Lee, B. H. Kim, and J. Seo. ManiWordle: Providing flexible control over wordle. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1190–1197, 2010.

[22] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. SparkClouds: Visualizing trends in tag clouds. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1182–1189, 2010.

[23] S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *Int. Work. Conf. Advanced Visual Interfaces*, AVI '12, pp. 753–756. ACM, 2012.

[24] S. Lohmann, F. Heimerl, F. Bopp, M. Burch, and T. Ertl. ConcentriCloud: Word cloud visualization for multiple text documents. In *19th Int. Conf. on Information Visualisation*, IV '15, pp. 114–120. IEEE, 2015.

[25] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In *12th IFIP TC 13 Int. Conf. on Human-Computer Interaction*, INTERACT '09, pp. 392–404. Springer, 2009.

[26] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[27] 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 *SIGCHI Conf. on Human Factors in Computing Systems*, CHI '07, pp. 995–998. ACM, 2007.

[28] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[29] J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, 2008.

[30] H. Strobelt, M. Spicker, A. Stoffel, D. A. Keim, and O. Deussen. Rolled-out wordles: A heuristic method for overlap removal of 2d data representatives. *Comput. Graph. Forum*, 31(3):1135–1144, 2012.

[31] Tagul. Gorgeous tag clouds. `http://tagul.com`.

[32] Tagxedo. Word cloud with styles. `http://www.tagxedo.com`.

[33] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1986.

[34] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: a site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1121–1128, 2007.

[35] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar. What's being said near "martha"? exploring name entities in literary text collections. In *IEEE Symp. on Visual Analytics Science and Technology*, VAST '09, pp. 107–114. IEEE, 2009.

[36] Y. Wang, X. Chu, C. Bao, L. Zhu, O. Deussen, B. Chen, and M. Sedlmair. EdWordle: Consistency-preserving word cloud editing. *IEEE Trans. Vis. Comput. Graph.*, 24(1):647–656, 2018.

[37] Wordle. Beautiful word clouds. `http://www.wordle.net`.

[38] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: interactive visualization of hotel customer feedback. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1109–1118, 2010.