ConcentriCloud: Word Cloud Visualization for Multiple Text Documents

Steffen Lohmann, Florian Heimerl, Fabian Bopp, Michael Burch, Thomas Ertl

Institute for Visualization and Interactive Systems (VIS), University of Stuttgart, Germany
{steffen.lohmann,florian.heimerl,michael.burch,thomas.ertl}@vis.uni-stuttgart.de

***Abstract*—Word clouds provide a simple and effective means to visually communicate the most frequent words of text documents. However, only few word cloud visualizations support the contrastive analysis of multiple texts. This paper introduces ConcentriCloud, a layered word cloud layout that merges the words from several text documents into a single visualization. The weighted words are arranged in a concentric layout, with those representing the individual documents on the outer circle and the merged ones on inner circles. Interaction techniques allow to analyze the word cloud composition and to provide details on demand. The approach has been implemented and tested on several examples. A qualitative evaluation indicates the general value of ConcentriCloud and reveals benefits and limitations.**

***Index Terms*—Text visualization, word cloud, tag cloud, text documents, information visualization, ConcentriCloud.**

I. Introduction

Word clouds are a simple and intuitive visualization technique that is often used to provide a first impression of text documents. Typically, they show the most frequent words of a text as a weighted list of words in some specific spatial layout (e.g., sequential, circular, random) [15]. The font sizes of the words indicate their relevance or occurrence frequency, while other visual properties (e.g., color, position, orientation) are often varied for aesthetic reasons or to visually encode additional information.

Word clouds can serve as a starting point for deeper text analyses [10], [24]. However, available word cloud visualizations provide only limited support in comparing the words and word frequencies of different text documents. To overcome this limitation, we propose ConcentriCloud, an extended word cloud visualization that systematically merges and displays the words from several text documents. It gives an overview of the documents and makes differences and commonalities in word use immediately visible.

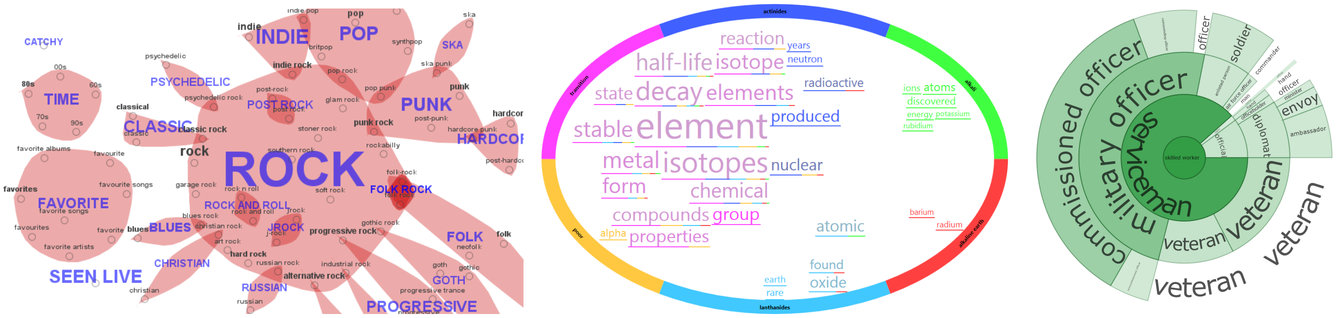
Basically, a ConcentriCloud is composed of a number of smaller word clouds that represent different combinations of the documents. The word clouds are arranged in a concentric layout, with those representing the individual documents on the outer circle and the merged ones on inner circles. The word cloud in the innermost circle contains the words that occur in all documents. This composition principle is emphasized by the saturation of the background color, which increases with the level of aggregation. Interaction techniques allow to further analyze the word cloud visualization and to provide details on demand.

This paper presents ConcentriCloud in detail. After summarizing the related work (Section II), we describe the visualization concept (Section III) and its implementation (Section IV). We show some possible interaction techniques and demonstrate the applicability the approach on selected examples. Finally, we report on a qualitative evaluation of ConcentriCloud (Section V) before we conclude the paper (Section VI).

II. Related Work

Several extensions to the basic word cloud visualization have been proposed in recent years. One line of research concerns the improvement of the layout of word clouds. For instance, Kaser and Lemire [11] present methods to reduce and balance the white space in word clouds using rectangular layouts. Seifert et al. [19] developed other algorithms for space-filling word clouds based on simple heuristics that can cope with polygonal shapes. Further layout algorithms have been proposed in the works on ManyWordle [12] and Rolled-out Wordles [21], among others. Advanced designs are also used in online word cloud generators, such as Wordle, Tagul, or Tagxedo. Although the general layout of our ConcentriCloud approach is determined by its concentric design, different layout strategies may be applied to distribute the weighted words in the individual word clouds the ConcentriCloud is composed of.

Some layout strategies consider word relationships (e.g., based on co-occurrences) and implement spatial arrangements where strongly related words are placed in close proximity. The layout strategies range from simple line-by-line approaches [9] to treemap-like layouts [11] and force-directed placements in combination with Venn diagrams (cf. Figure 1a) [4]. Other works apply 2D projection techniques based on multidimensional scaling to reflect the relatedness of words [17], [25], or use topographical word landscapes [8]. There are also attempts to explicitly depict relationships in word clouds, either by adding links between related words [20] or by using interactive highlighting [10], [14]. Prefix Tag Clouds [2] make use of prefix trees to group different word forms, whereas the Word Cloud Explorer uses advanced NLP processing to link word forms and to support the visual analysis of text documents via interactive word clouds [10].



(a) TagCluster (b) RadCloud (c) DocuBurst

Fig. 1. Selection of related visualizations.

However, all these works do not distinguish different text documents, but the words are treated as if they would all come from the same text source, even if they do not. This is also true for approaches that add a temporal dimension to word clouds, for instance, by using sparklines [13] or histograms [14] in order to indicate changes in word use over time. While these visualizations can be used to illustrate the evolution of words in different text documents, the documents themselves are not distinguished in the word clouds. The same limitation holds for the work of Cui et al. [7] who coupled trend charts with word clouds to illustrate the temporal evolution of words.

More related to our approach are works that use small multiples of word clouds to visualize several documents at a time. OpinionSeer [26] implements this attempt to visually summarize and compare the reviews of hotel customers, while FacetScape [18] arranges multiple word clouds in a Voronoi diagram. This multiplication approach can also be found in POSvis [24], where each small word cloud is dedicated to a specific part-of-speech category. The Word Storms approach [3] even makes sure that words appearing in multiple documents are placed in the same location across clouds, with the same color and orientation, to improve the comparability. However, the same words are displayed multiple times in all these approaches, and it still remains difficult to compare individual words and their frequencies, as the word clouds for the different texts are not explicitly linked but only implicitly by color and style.

A similar limitation is given in Parallel Tag Clouds [6] that combine the ideas of word clouds, small multiples, and parallel coordinates by making each coordinate a weighted list of words. Words once again appear multiple times, and interaction is required to understand how they are linked and used in the different text sources. An exception in this regard is the word cloud visualization of the ManyEyes website [23] that shows words from more than one text in a single word cloud, using font color to indicate the text source of each word. However, also in this approach, words appear multiple times, and it does not scale well if more than two or three texts are visually compared.

DocuBurst [5] uses a sunburst visualization to show a hierarchy of concepts extracted from text documents (cf. Figure 1c). Apart from the fact that different texts are again not distinguished in Docuburst, such hierarchical approaches are different from ConcentriCloud, which follows a set-theoretic approach instead when combining the word clouds of the texts. This is also the reason why other related work on the hierarchical composition of word clouds, such as the treemap-based approach of the ScatterScopes system [22], is different from ConcentriCloud.

Most closely related to the idea of ConcentriCloud is the RadCloud approach [1]. It also merges words from multiple text sources and visualizes them in a circular word cloud, with the individual text documents forming the circle border (cf. Figure 1b). However, instead of a concentric design that clearly indicates the composition of the visualization, RadCloud uses a relaxed force-directed layout. This layout tends to place words in quite some distance from the originally computed positions. Furthermore, an efficient use of screen space is computationally expensive with the RadCloud algorithm and would require some clever heuristic strategies. This is different in ConcentriCloud where we aim for a space-filling approach that clearly communicates the composition of the visualization and can be computed in comparatively little time.

III. ConcentriCloud

Basically, a ConcentriCloud is composed of several word clouds representing different combinations of the text documents. This is akin to the small multiples idea as it offers comparable views of the document terms. However, through a systematic merging of the word clouds, commonalities and differences between the documents can be more easily identified. Furthermore, the merging avoids redundancies, as terms are usually displayed only once in ConcentriCloud (with the exception of some special cases that are later discussed).

A. Composition Principle

ConcentriCloud arranges the word clouds on concentric circles, as sketched in Figure 2 for four documents A,B,C, and D. Each document is represented by a *bag of words* that comprises all terms from the document (or a cleaned subset of terms) along with their frequency values. Formally, each document can therefore be defined as a set of terms

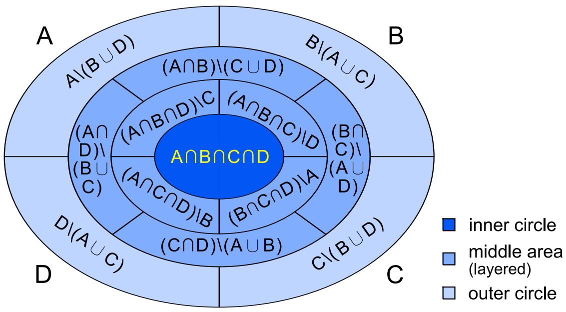


Fig. 2. Schematic illustration of the composition of ConcentriCloud (letters A to D represent the bags of words of four text documents).

$T := \{t_1, \ldots, t_n\}$ with individual terms t_i, $1 \leq i \leq n$. Each term is additionally associated with a quantitative value expressing its occurrence frequency. The sets of terms T_x are combined into several word clouds W_y, each representing a different combination of the documents.

The word clouds on the outermost circle contain terms that occur in either of the documents (with the below exception). For instance, the word cloud representing document A consists of terms that are contained in A but not in B and C (i.e., $W_A = A \setminus (B \cup C)$). Note that document D is an exception in this example, as its word cloud is located at the opposite side of the outer circle. Opposite word clouds are *not* merged on any layer of the middle area, as there is no intuitive position for such a composite cloud in the concentric layout, except from the inner circle. The inner circle, however, is reserved for words that occur in all documents and not only in a specific subset of documents. This results in the fact that terms can appear more than once in the visualization if two opposite word clouds contain the same term. However, this redundancy due to layout constraints only marginally affects the general readability and interpretation of the visualization, as we found in the qualitative evaluation (cf. Section V)—especially if it is visually indicated, for instance, by interactive highlighting.

In each of the word clouds towards the center, the terms from the documents are systematically combined, i.e., the layers on the middle area contain terms that occur in more than one document (but *not* in all documents). For instance, the next inner circle contains word clouds that represent the pairwise intersections of the documents minus the pairwise unions of the rest of the documents (with the aforementioned exception that oppositely located word clouds are not combined). In case of documents A and B, this results in a word cloud $W_{A,B} = (A \cap B) \setminus (C \cup D)$, among others (cf. Figure 2). Finally, the innermost circle consists of only one word cloud containing those terms that occur in *all* documents. In the illustrated case, it thus represents the intersection of all four documents, i.e., $W_{A,B,C,D} = (A \cap B \cap C \cap D)$.

B. Further Considerations

For any other number of documents, the composition of the visualization needs to be adapted accordingly. In case of three documents, the ConcentriCloud consists of three circles; in case of five documents, it consists of five. As a general rule, each ConcentriCloud is theoretically composed of as many circles as there are documents. Since this can result in a large number of circles, certain layers in the middle area may be skipped, as long as the overall composition principle remains the same. However, note that terms should always appear on the highest possible aggregation level in ConcentriCloud, i.e., on the level closest to the center.

The only exception are terms from oppositely located word clouds, or, more generally, from word clouds that are not neighbors on the outer circle. In case of three documents, there is no such exceptional case, but with an increasing number of documents, the likelihood of term redundancy increases, as not all documents can be combined in the layout. One strategy to minimize any remaining term redundancy in the visualization is to order the documents based on term similarity, as we have implemented it in our prototype (cf. Section IV).

C. Visualization Example

Figure 3 shows an example of a ConcentriCloud visualizing frequent terms of all seven "Harry Potter" novels. The word clouds on the outermost circle represent the individual novels (HP1 to HP7). They are visually separated by lines, while the names of the source files are shown next to them. Examples of terms that appear only in *one* of the novels are "lockhart" (second novel) and "karkaroff" (fourth novel). The angular size of the word clouds indicates the relative length of each novel, which is increasing in the case of Harry Potter. Terms that can be found in all seven novels are shown in the inner circle of the visualization, such as "harry" or "dumbledore".

Since the inner and outer circle are most important for the idea of ConcentriCloud, they are clearly distinguished in the visualization. Borders between the layers in the middle area are omitted to produce a clearer picture and to reduce visual clutter that would be introduced by too many separating lines. If a word cloud in the middle area does not require all the reserved screen space, it is used by neighboring word clouds to place further terms beyond their bounding box for a more space-filling design. However, the general composition principle remains the same, i.e., the closer a term to the center the more documents contain it. This principle is additionally emphasized by the saturation of the background color, which has a gradient towards the center in the middle area.

IV. Implementation

We implemented the approach in a Java-based prototype that generates ConcentriClouds as the one presented in Figure 3. In the following, we describe key issues and design decisions related to the implementation.[1]

A. Text Processing

The first step in creating a ConcentriCloud is to process the text documents and extract meaningful terms and their occurrence frequencies. For this task, we use the Stanford CoreNLP framework [16] that includes standard NLP pipeline

[1]A demo video is available at http://wordclouds.visualdataweb.org.

Fig. 3. ConcentriCloud visualization of all seven "Harry Potter" novels (HP1 to HP7).

functionalities. In particular, we run the CoreNLP tools for tokenization, lemmatization, and part-of-speech tagging to process the text documents. The tokenizer splits the document string into a list of individual tokens. Those are processed by a lemmatizer and a part-of-speech tagger, labeling each token with its respective part-of-speech, and identifying its grammatical base form. Part-of-speech information can, for instance, be used to create word clouds that include only nouns, such as the Harry Potter word cloud in Figure 3. Identifying the lemma of each token can produce cleaner word clouds, as all morphological variations of a word, such as plural forms for nouns or conjugations for verbs, can be merged. The extracted set of lemmas is then filtered using a list of stopwords that contains high-frequency words which, in isolation, do not contribute any relevant information about the document's content (e.g., words like "the" or "and"). Finally, the terms are all converted to lowercase and transformed into the aforementioned bag-of-words representation required for the ConcentriCloud.

B. Concentric Layout

There are several possibilities of arranging the documents on the outermost circle. We have implemented two different approaches in our ConcentriCloud prototype: The first allows for a manual ordering by the user, which may reproduce some natural ordering of the documents, e.g., according to their publication date. An example for this are the Harry Potter novels in Figure 3, ordered from the earliest to the latest publication in this series. As a second way of arranging the documents, we developed an algorithm that orders them according to their similarities. After computing the cosine similarity for each pair of documents, the algorithm greedily chooses the highest similarity score between two documents and reduces the set of possible orderings to those in which both documents are neighbors. This is done recursively until each document has its fixed position. This similarity-based ordering algorithm was, for instance, used to create the ConcentriCloud visualization of patent documents shown in Figure 4.

After the ordering of the documents, the size of the word clouds on the outer circle is determined by scaling the angle according to the document length. We implemented a scheme that is able to accommodate the worst case scenario in which one very long document will take all the space from several very short documents. This is solved by splitting the available $360°$ into two parts, assigning each of the n documents a minimum angle of $\frac{180°}{n}$, and allotting the remaining $180°$ degrees according to document size.

ConcentriCloud attempts to render the terms within the bounding box of the respective word clouds between the concentric circles. We used a layout strategy similar to the one described in [2] in our ConcentriCloud implementation: The terms are placed along invisible concentric circles from the center of the respective word cloud, starting with the most frequent term and continuing with terms of decreasing frequencies. This ensures that the most frequent terms are tried to be placed first and that they appear in the center of the respective word cloud, such as the term "harry" in the inner

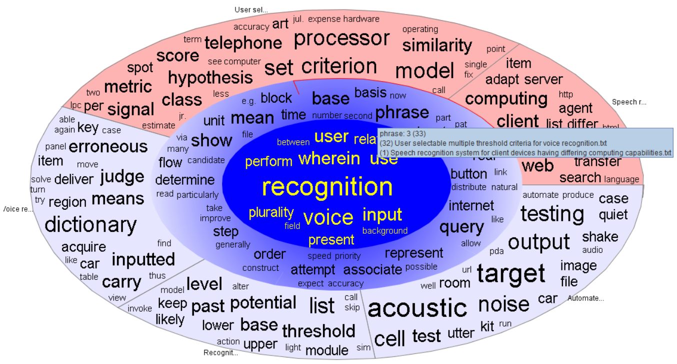


Fig. 4. ConcentriCloud visualization of five patents on the topic "voice recognition" (the term "phrase" is hovered and related patents are highlighted).

Fig. 5. Reference visualization for the "Harry Potter" use case.

cloud of Figure 3. If a term cannot be placed within the cloud's bounding box, it is omitted from the word cloud, and it is tried to place the next term from the frequency-ordered list of terms into the cloud.

This placement strategy has the limitation that some high-frequency terms may not be rendered due to their larger size and space limitations, whereas other low-frequency terms of smaller size could be placed in the word cloud, as they fit in the available space. However, this is rather a general limitation of word clouds than a particular drawback of ConcentriCloud. An alternative strategy would be to stop the placement and do not add smaller terms in the available space, as soon as a larger term cannot be placed. Yet, this could result in a lot of unused space that may better be filled with smaller terms, also for aesthetic reasons.

The font size of the terms is scaled either linearly or logarithmically with their occurrence frequency, depending on the overall frequency distribution in the documents. If a word cloud represents the terms from more than one document, we use the average of the term frequencies for scaling the font size.

C. Interaction

Our implementation does not only create a static ConcentriCloud, but it includes options to customize and interact with the visualization. Before the visualization is created, users can specify the ordering of the documents, and if they would like to include all terms or only nouns in the word clouds. As already mentioned, the noun-only mode can help to produce more lucid word clouds, depending on the analyzed document set. While Figure 3 was an example of a noun-only ConcentriCloud, Figure 4 displays also other parts of speech, such as verbs and adjectives.

Figure 4 also shows two further modes of interaction: (1) highlighting of the word clouds that represent the corresponding documents when the user hovers a term in a cloud (in this case "phrase"), and (2) tooltips that appear for each word showing its overall number of occurrences in the document set and its distribution across the individual documents. These interaction modes assist the user in getting a better understanding of the composition of the ConcentriCloud and the exact term frequencies.

V. Evaluation

We conducted a qualitative evaluation with expert users to get feedback on the design of our approach, its strengths and weaknesses, as well as implementation-specific issues. For this, we recruited a total of six researchers from the visualization department of our university, one female and five males between the age of 27 and 32 years. None of the expert users has come in contact with the ConcentriCloud approach before or was biased with regard to the visualizations presented in the study.

A. Design

To get deeper insights into the properties and analytical capabilities of the circular design we chose for our approach, we devised and implemented an alternative layout for the word clouds. As depicted in Figure 5, the alternative layout was rectangular and the word clouds are arranged from top (individual documents) to bottom (entire document set). The bags of words used to generate this layout were identical to those in the circular design, only the bounding boxes of the word clouds were transformed into a rectangular shape. The color coding as well as the interaction possibilities on the alternative layout were exactly the same as on the original design.

We presented both layouts to each of the experts on a 15.6 inch screen with a resolution of 1366×768 pixels. Before we introduced the implementation to the experts, we tested them for color vision deficiencies using the Ishihara color plates. In addition, we asked them to judge their previous knowledge about word cloud visualizations, text analysis, and the Harry Potter novel series, each on a scale of 1 to 10. We were asking for Harry Potter, as we were using the seven novels as an evaluation dataset (cf. Figure 3). Apart from the Harry Potter texts, we created a training dataset containing patent documents and visualized it as ConcentriCloud (cf. Figure 4). It comprises five patents that deal with voice recognition technology.

TABLE I
QUESTIONS OF THE USER STUDY.

Questions and answers for the ConcentriCloud visualization shown in Figure 3.	
Which term does occur in every novel?	(Harry)
Which term does occur most often only in the 3rd novel?	(pettigrew)
To which novel does the term "elf" belong?	(4,5,6,7)
Which term is the most frequent on average in the first and second novel?	(Justin)
How often does "Harry" occur in the first novel?	(1306)
Questions and answers for the reference visualization shown in Figure 5.	
Which term does occur least often in all novels?	(fighting)
In which of the novels does the term "griphook" occur?	(1,7)
Which term has the highest frequency of those that only occur in the second novel?	(lockhart)
Which term does on average occur most often in novels 3, 4, and 5?	(boggart)
How often does "Ron" occur in the fourth novel?	(1042)

The experts were introduced to both visualizations of the *voice recognition* dataset and asked to answer three simple questions about the dataset to learn how to interpret the layouts and how to use the implementation. After that, we presented the two visualizations of the Harry Potter novels to the experts, i.e., the ConcentriCloud layout and the alternative layout. To counterbalance any effects introduced by the presentation order of the layouts, we were using each of the two possible orders with half of the experts. We asked each participant to answer five questions about the Harry Potter novels for each of the layouts. English translations of all ten questions are listed in Table I. We designed the questions to test how well the participants are able to read and interpret the visualizations.

During the study, we were using the think-aloud method, encouraging the participants to ask questions and give feedback at any time. In the subsequent interview, we were asking for positive and negative impressions of the general approach and the two layouts. Additionally, the experts were asked to pick their favorite layout and to elaborate on the reasons for their choice. Finally, we were collecting any remarks on problems or bugs of the implementation and any functionality that the experts were missing.

B. Results

All six participants passed the Ishihara test without any color vision deficiencies detected. On a scale of 1 to 10, the average knowledge about word clouds was 5.0 (min: 2, max: 8, SD: 2.3), the proficiency in text analysis was rated a little higher with an average of 5.7 (min: 2, max: 8, SD: 2.1). Harry Potter seems to be popular among the experts, scoring an average of 6.2 (min: 2, max: 9, SD: 2.7).

All participants were struggling with the interpretation of the center cloud during the initial questions, but they learned how to read it within a few seconds up to one minute time. They were readily able to effectively navigate the different clouds once they figures out their interpretation. All questions could be answered correctly at the end.

Several participants were missing a search function for specific terms, as they were spending quite some time to visually search for terms. They also mentioned that changing the ordering of the terms in the clouds to an alphabetical one could be helpful to find a specific term more quickly. Some participants were missing a feature that shows selected terms in their text context, as some terms (e.g., unusual person names) were hard to interpret in isolation. They also mentioned that it might help to show co-references for each name from the text. Due to the visual blending of multiple word clouds on the middle area, some users considered it difficult to find terms for a specific combination of documents on this level. One suggested an extension that lets users mark any number of documents, and subsequently highlights the terms occurring in all of them. Nevertheless, all participants found that the layout of terms on the middle area is coherent and can generally be interpreted correctly.

The participants unanimously mentioned the missing highlighting of redundant occurrences of terms in word clouds as a problem. As mentioned, redundancy occurs in ConcentriCloud if a term is part of word clouds of non-neighboring documents. Another problem, addressed by a single participant, is the possibility of one document dominating a word cloud in the middle area with respect to one specific term. If there are, for instance, three documents that contain "elf", one contains it 100 times, and each of the others contain it once, the term will prominently show up in the combined cloud for the three documents instead of in the cloud of the document that has over 98% of its occurrences. A way of solving this problem would be to define a threshold for the maximum frequency ratio for a term that two documents may have to include it on the middle level.

With the alternative layout, the participants found it much less intuitive to link the terms on the middle level to their corresponding documents. This is due to the offset between the document positions and the corresponding clouds in the middle area. The layout, however, has the positive property of being able to accommodate more terms within the clouds, because the rectangular space is used more efficiently. Some participants, on the other hand, found that this characteristic of the alternative layout quickly leads to cluttered word clouds with too many terms. The radial layout was praised for its compact representation of a document set and its lucid depiction of common terms of documents. Compared to the alternative layout, all participants preferred the radial one. After a brief learning phase, all users were able to interpret the ConcentriCloud visualization, and to use it effectively. At the same time, the radial layout was rated as being the aesthetically more pleasing and clearer depiction of the text documents.

VI. CONCLUSION

We have presented ConcentriCloud, a novel visualization based on word clouds that systematically merges the terms from several text documents. The word clouds are arranged in a concentric layout, with those representing the individual documents on the outermost circle and the merged ones on inner circles. ConcentriCloud provides a first impression of the word use in the documents and supports the visual identification of differences and commonalities. Interaction techniques allow to further analyze the visualization and to provide details on demand. The approach has been implemented and tested on several examples, and a user study has been conducted that confirms its general value.

In principle, the ConcentriCloud visualization can scale up to an arbitrary number of documents and words, but it would usually not make much sense to visualize the words from more than a handful or maybe a dozen of documents, as this would become too demanding for the viewer. Moreover, it is important to note that word clouds usually do not show all terms of a text document but only the most frequent ones. Due to the space-filling layout, smaller words may be added to the word cloud if larger ones do not fit in the remaining screen space. In order to avoid a wrong interpretation in these cases, we recommend to provide a list of the actual terms and term frequencies for each word cloud on demand.

Furthermore, there are some extreme cases one should be aware of when using ConcentriCloud. For instance, if the analyzed documents do not have any single word in common, the inner circles would be empty and only the word clouds on the outermost circle would display terms. In the opposite case of text documents that share (nearly) all words, the outer circle would be empty and terms would only appear in the word clouds of the inner circles. Although such extreme cases are rather unlikely, they illustrate the limitations of the approach and indicate that it may not work equally well in all situations. Despite these limitations, we believe that the general approach has much potential, especially as there are only very few works that address the problem of combining multiple documents in a single word cloud visualization (cf. Section II).

The qualitative evaluation revealed some issues that may be addressed in future work, such as the optimal placement of words in the middle area. Another direction of research could be the integration of ConcentriCloud with related attempts, such as the RadCloud approach [1] or the Word Cloud Explorer [10] (cf. Section II). In particular, additional interactive features could be added that extend the analytical capabilities of ConcentriClouds, such as a term search, the highlighting of term relations, as well as possibilities to easily look up words in the original text context. However, such extensions are beyond the scope of this paper and independent from the main contribution of ConcentriCloud.

REFERENCES

[1] M. Burch, S. Lohmann, F. Beck, N. Rodriguez, L. D. Silvestro, and D. Weiskopf. RadCloud: Visualizing multiple texts with merged word clouds. In *18th International Conference on Information Visualisation*, IV '14, pages 108–113. IEEE, 2014.

[2] M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf. Prefix tag clouds. In *17th International Conference on Information Visualisation*, IV '13, pages 45–50. IEEE, 2013.

[3] Q. Castellà and C. Sutton. Word storms: Multiples of word clouds for visual comparison of documents. In *23rd International Conference on World Wide Web*, WWW '14, pages 665–676. ACM, 2014.

[4] Y.-X. Chen, R. Santamaría, A. Butz, and R. Therón. TagClusters: Semantic aggregation of collaborative tags beyond TagClouds. In *10th International Symposium on Smart Graphics*, SG '09, pages 56–67. Springer, 2009.

[5] C. Collins, M. S. T. Carpendale, and G. Penn. Docuburst: Visualizing document content using language structure. *Computer Graphics Forum*, 28(3):1039–1046, 2009.

[6] C. Collins, F. B. Viégas, and M. Wattenberg. Parallel Tag Clouds to explore and analyze faceted text corpora. In *IEEE Symposium on Visual Analytics Science and Technology*, VAST '09, pages 91–98. IEEE, 2009.

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

[8] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda. Topigraphy: visualization for large-scale tag clouds. In *International Conference on World Wide Web*, WWW '08, pages 1087–1088, 2008.

[9] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*, InSciT '06, pages 25–28, 2006.

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

[11] 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.

[12] K. Koh, B. Lee, B. Kim, and J. Seo. ManiWordle: Providing flexible control over wordle. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1190–1197, 2010.

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

[14] S. Lohmann, M. Burch, H. Schmauder, and D. Weiskopf. Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds. In *International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 753–756. ACM, 2012.

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

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, ACL '14, pages 55–60. ACL, 2014.

[17] F. V. Paulovich, F. M. B. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato. Semantic wordification of document collections. *Computer Graphics Forum*, 31(3):1145–1153, 2012.

[18] C. Seifert, J. Jurgovsky, and M. Granitzer. FacetScape: A visualization for exploring the search space. In *18th International Conference on Information Visualisation*, IV '14, pages 94–101. IEEE, 2014.

[19] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer. On the beauty and usability of tag clouds. In *12th International Conference on Information Visualisation*, IV '08, pages 17–25, 2008.

[20] M. Stefaner. Visual tools for the socio-semantic web. Master thesis, University of Applied Sciences Potsdam, 2007.

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

[22] D. Thom, M. Wörner, and S. Koch. Scatterscopes: Understanding events in real-time through spatiotemporal indication and hierarchical drilldown. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 387–388, 2014.

[23] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. ManyEyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.

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

[25] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-preserving word clouds by seam carving. *Computer Graphics Forum*, 30(3):741–750, 2011.

[26] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: Interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization on Computer Graphics*, 16(6):1109–1118, 2010.