Word Cloud Explorer: Text Analytics based on Word Clouds

Florian Heimerl, Steffen Lohmann, Simon Lange, Thomas Ertl
Institute for Visualization and Interactive Systems (VIS)
University of Stuttgart, Germany

Abstract

Word clouds have emerged as a straightforward and visually appealing visualization method for text. They are used in various contexts as a means to provide an overview by distilling text down to those words that appear with highest frequency. Typically, this is done in a static way as pure text summarization. We think, however, that there is a larger potential to this simple yet powerful visualization paradigm in text analytics. In this work, we explore the usefulness of word clouds for general text analysis tasks. We developed a prototypical system called the Word Cloud Explorer that relies entirely on word clouds as a visualization method. It equips them with advanced natural language processing, sophisticated interaction techniques, and context information. We show how this approach can be effectively used to solve text analysis tasks and evaluate it in a qualitative user study.

1 Introduction

Having their roots "outside the world of computers" [34], tag clouds became popular in the context of community-oriented websites, such as Flickr, Delicious, or Technorati, that use tagging as an indexing method [27]. Meanwhile, they have evolved as a core technique of information visualization that is applied in many different contexts.

One popular application area for tag clouds is text summarization [2, 6, 15]. Here, tag clouds are used to give an intuitive and visually appealing overview of a text by depicting the words that occur most often within it. Such a summarization is helpful to learn about the number and kind of topics present in a body of text. Typically, this statistical overview is achieved by positively correlating the font size of the depicted tags with the word frequency. When a tag cloud visualization is used this way, the 'tags' are words from a text. For this reason, the term *word cloud* is often preferred over the term *tag cloud* in these contexts. We will also use it in the remainder of this paper.

Word clouds generated for a body of text can serve as a starting point for a deeper analysis [2, 26, 35]. For instance, they help to judge whether a given text is relevant to a specific information need. One of their drawbacks is that they provide a purely statistical summary of isolated words without taking linguistic knowledge about the words and their relations into account. Consequently, word clouds are used rather statically as a means to summarize text in most systems and they typically provide no or only limited interaction capabilities.

We think there is a larger potential to this simple yet powerful visualization paradigm in many analysis contexts. In this work, we therefore explore the possibilities of using word clouds at the very center of text analysis. We developed the *Word Cloud Explorer*, a prototypical system that uses word clouds as its main visualization and interaction hub. We equipped it with advanced natural language processing, sophisticated interaction possibilities, and a high level of control for users to provide support for different kinds of text analysis tasks. Users can drill down to the local contexts of words and use flexible filter mechanism in combination with linguistic information for further analysis. The intuitiveness of the word cloud visualization makes the Word Cloud Explorer a system that is easy to learn.

The main contributions of this work are: *i)* a text analytics approach based on word clouds that offers a broad range of interactive and analytical features; *ii)* an easy to use yet powerful and highly configurable implementation of the approach, demonstrating the feasibility and effectiveness of the approach; *iii)* a qualitative user study that yields further insight into the approach.

2 State of the Art

Research on word clouds falls in one of two categories: 1) work that investigates the effectiveness and visual perception of word clouds, and 2) work that develops improvements and extensions to the word cloud visualization. In addition, we define a third category for this work, consisting of text analysis systems that use word clouds as one of their components.

To appear in: Proceedings of the 47th Hawaii International Conference on System Science (HICSS 2014), January 6-9, 2014
IEEE Computer Society

2.1 Effectiveness and Perception

There have been several attempts to investigate the effectiveness and perception of word clouds. Bateman et al. [1] conducted a user study in which they systematically varied nine visual properties of word clouds. They found that the properties with the largest effect on the users' attention are font size, weight, and color. Rivadeneira et al. [22] also observed a strong effect of font size in their user study. Furthermore, Bateman et al. as well as Lohmann et al. found that terms in the middle of the cloud receive more attention on average than terms near the borders [1, 19].

Word clouds have been compared to unweighted lists and other user interfaces in a number of studies [10, 19, 22]. The results indicate that users are on average more effective in spotting a specific term in an alphabetically ordered unweighted list than in an alphabetically ordered word cloud. However, frequently used terms are found more quickly in word clouds due to their larger font sizes [1, 19, 22].

Sinclair and Cardew-Hall [26] compared word clouds with a user interface simply consisting of a search box. While participants preferred the search box to enter specific terms, they favored the word cloud for more open-ended tasks. This finding is supported by Kuo et al. [15] who used word clouds to summarize search results. Their results further indicate that word clouds are effective to give an impression of what information is present in a query result set. They draw the conclusion that word clouds are a good visualization technique to communicate an 'overall picture' of the text contents.

2.2 Improvements and Extensions

The above reported works mainly studied rectangular word clouds with a sequential line-by-line layout. However, several improvements and extensions to this basic layout have been proposed in the last couple of years. Kaser and Lemire [13], for instance, use slicing trees, nested tables, and rectangle packing to optimize the distribution of space in HTML-based word clouds. Seifert et al. [25] present a related algorithm for white space optimization that can cope with differently shaped word clouds. It places terms in a circular fashion with the most frequent ones at the center and those with lower frequency towards the boundaries.

Other works use clustering techniques along with different kinds of word cloud layouts, ranging from line-by-line layouts [11, 24] to force-directed layouts [3] and topographical term landscapes [8]. While the relatedness of the terms is indicated by their spatial distance in most of these works, some explicitly depict term relations, either by connecting the terms with arcs [30] or by highlighting related terms in the word cloud [5, 18].

Over the last years freely available word cloud generators, such as Wordle [37], Tagul [31], or Tagxedo [32], have been developed that produce visually appealing word clouds. These tools offer several options to customize the word cloud visualization by adapting typography, color, word orientation, or even the general shape of the word cloud. However, they are intended as 'design tools' rather than tools for text analytics. Consequently, the resulting word clouds are aesthetically pleasant but provide nearly no features to analyze the underlying text.

There are also attempts to include a temporal dimension in word clouds, for instance by using sparklines [16] or histograms [18] to depict changes in term use over time. Parallel Tag Clouds [4] combine the ideas of word clouds and parallel coordinates to allow for a direct comparison of term frequencies at different points in time or from different data sources. Tree Clouds [9] combine word clouds with trees to visualize the semantic relatedness of terms. Prefix Tag Clouds [2] use prefix trees to group different word forms and visualize the subtrees as tag clouds. Finally, there are also 3D variants of word clouds, such as WP-Cumulus [38] that provides a rotating, three-dimensional sphere of terms.

While a part of these extensions has been designed for specific application contexts, others can be used more generically. We adopted some of these ideas in our approach, such as the circular word cloud layout or the interactive highlighting of term relations.

2.3 Applications in Text Analysis

There are several text analysis systems that make use of word clouds. Examples can be found in domains such as patent analysis [14], opinion mining [39], or investigative analysis [29]. In most of these systems, word clouds are used in a static way to visually summarize text documents.

An interactive word cloud variant has been implemented in the VisGets system [5]. It is used to filter terms in web-based information retrieval. Multiple terms can be concurrently selected as filters, restricting the items in the result list to only those including these terms. Related terms in the word cloud are highlighted, as well as related elements in the other views, consisting of a temporal bar chart and a geographical map. However, the word cloud is used as just one visualization component among many in the VisGet system. Although it is connected with other views by brushing and linking, the analytical possibilities of the word cloud itself are limited.

A noteworthy exception is the POSvis system [36]. It uses word clouds as one of its main views and is therefore closely related to our work. POSvis is a lit-

erary analysis system helping scholars to review the vocabulary of novels, filter it by parts of speech, and explore networks of characters from the novel. It supports these tasks with two main word cloud views: One that displays extracted character names and another showing words from selected part-of-speech categories. In addition, it offers a third view consisting of multiple word clouds, each dedicated to a specific part-of-speech category. Users can customize the order, size, and color of the words in the cloud.

However, the analytical and interactive features of the word clouds in the POSvis system are limited compared to our approach. Our Word Cloud Explorer offers generic focus-and-context techniques and direct interaction with the word cloud. It provides a broad range of interactive features and is highly configurable. This makes it suitable for a wide variety of text analytics problems, while POSvis has mainly been designed for the analysis of novel characters and their relationships.

3 The Word Cloud Explorer

Figure 1 shows a screenshot of the Word Cloud Explorer, generated with the popular Sherlock Holmes novel "The Hound of the Baskervilles" by Arthur Conan Doyle. The system consists of the central word cloud view and a number of other components providing additional information and functionality for the analysis. The individual components are marked with letters in Figure 1. In the following, we will describe their functionality and explain how they support analysts.[1]

3.1 Text Processing

After a text file is loaded, the system performs a linguistic analysis of its contents. We use the Stanford CoreNLP tools [28] for this purpose and perform several processing steps, consisting of tokenization, sentence splitting, part-of-speech tagging, lemmatization, and named-entity recognition. Based on the results of the part-of-speech tagger, we additionally implemented a detector for nominal multiword expressions. It joins all continuous sequences of proper nouns that occur in the same sentence. With this simple heuristic, we can detect most compound nominals and proper names in the text (see [23] for a comprehensive summary of different multiword phenomena).

The separate display of multiword expressions is important for many analysis tasks, especially those involving the identification of person or place names which are often multiwords (e.g. "Michael Jordan" or "New York"). Another benefit of considering multiwords is that the frequency counts of the individual terms are not artificially increased (e.g. "new" as part of "New York").

[1] Additional information on the Word Cloud Explorer, including an example video, can be found at http://wordclouds.visualdataweb.org.

3.2 Word Cloud View

The word cloud view (Figure 1ⓐ) is the main view of the system. It implements three different word cloud layouts that the users can choose from: two sequential line-by-line layouts, one ordered alphabetically, the other by frequency, and a circular layout showing the terms with the highest frequency in the center of the cloud and the lower frequency terms closer to the perimeter. The circular layout is shown in the screenshot of Figure 1; similar layouts have been presented in [2] and [25].

While the alphabetical layout supports users in quickly spotting specific terms they are looking for, the frequency ordered layout lets them arrange terms according to how often they occur in the text. The circular layout complements the sequential line-by-line layouts as a space-efficient and visually appealing alternative. Circular layouts have additionally shown to be most effective to spot high frequency terms in word clouds [19]. Font size is scaled linearly with the occurrence frequency of the terms for all layouts. As the Word Cloud Explorer allows for a free placement of terms, additional word cloud layouts can easily be added to the system. Also, the mapping of the frequencies to the font size of the terms may be adapted.

The word cloud view uses the information about different word forms provided by the lemmatization component to subsume them under one representative term in the word cloud. This means that, for example, not all inflections of a verb are shown as separate terms in the cloud but that their counts are added up and contribute to their most frequent representative within the given text. Detected multiwords are displayed in camel case to make them easily recognizable as one entity. However, both features can be disabled in the menu if users do not want a special treatment of multiwords or the merging of different word forms.

3.3 Co-occurrence Highlighting

Users can hover over terms to highlight related terms. In the current implementation, two terms are related if they co-occur within the same sentence. Alternative implementations might compute the co-occurrences on larger text segments, such as paragraphs or whole documents.

This feature of "co-occurrence highlighting" [18] is an intuitive technique to show term relations in word clouds without producing visual clutter. Related terms are marked with a yellow box in our approach whose saturation corresponds to the relative co-occurrence frequency. We have chosen this highlighter metaphor as it

Figure 1: The Word Cloud Explorer consists of the following components: ⓐ central word cloud view, ⓑ term filter, ⓒ search box, ⓓ term statistics panel, ⓔ info panel, ⓕ part-of-speech and named entity filters, ⓖ text viewer, ⓗ stopword editor, and ⓘ cloud control panel.

is very intuitive and provides an effective means to assess the 'strength' of term relations.

The second effect of hovering over a term is that further information about it is displayed in the term statistics panel ⓓ and info panel ⓔ we will describe in the following.

3.4 Term Statistics Panel

The term statistics panel ⓓ displays information about a focused term. This is illustrated for the term "looked" in Figure 1. The statistics panel lists the number of occurrences of the term within the filtered set of sentences (*Frequency filtered*), the number of terms currently present in the word cloud (*Terms filtered*), the number of occurrences of the selected term in the whole text corpus (*Frequency overall*), and the overall number of terms in the corpus (*Terms overall*). Finally, it gives the total number of sentences in which the focused term occurs (which is identical to the above values, as no filters are selected in this case).

The information from the term statistics panel can, for instance, be used for sanity checks. One apparent disadvantage of word clouds is that the difference in frequency between terms as judged according to their font size can give a false impression about the true frequency count ratio of the terms. Showing the absolute frequency values to users lets them easily correct false impressions.

3.5 Info Panel

The info panel ⓔ is part of the tabbed pane below the term statistics panel. It displays linguistic information about the focused term, consisting of the detected word

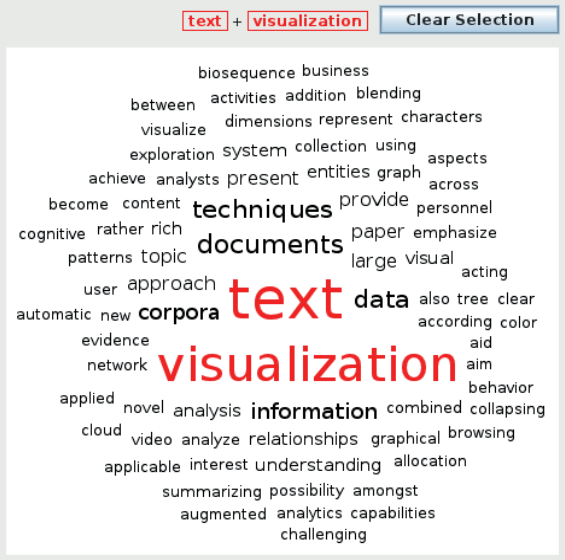


Figure 2: The terms *text* and *visualization* have been added to the term filter and are highlighted in the resulting co-occurrence cloud.

forms, part-of-speech tags, and named entity types for that term, along with the respective frequency counts.

In the example of Figure 1, five different word forms have been detected for the focused term. The word form "looked" is chosen as the representative because it appears most often in the text (as given by its count value). Furthermore, it has been detected that the term is mostly used as a verb in the text and that it is not a named entity (indicated by the category name "OTHER" in the list of named entities, see below).

This information helps analysts in two respects: First, it can be used to learn more about a term in the word cloud. For instance, it might be relevant for an analysis task if a term occurs mainly in present or past tense, or if it is used as verb or noun. Second, it can be used to cross-check the results of the linguistic analysis. Although the accuracy of the used text processing techniques is generally high, they are sometimes prone to errors depending on the type and quality of the text.

3.6 Term Filter and Search

Terms can be selected by clicking on them. They are then added to the term filter ⓑ. The word cloud view displays only those terms that co-occur with *all* of the selected terms, i.e. it changes from a word cloud for the whole text to a 'co-occurrence cloud' as soon as terms are selected. Figure 2 shows such a co-occurrence cloud for a text corpus with research abstracts from the field of visualization (see Section 5 for more details on this corpus). The selected terms are colored red (in this case,

Figure 3: A word cloud for the text "The Hound of the Baskervilles" with the terms colored according to their most frequent part-of-speech (noun, verb, adjective, adverb, preposition, other).

"text" and "visualization"). They can be added and removed from the term filter in any order and at any time, which triggers an update of the word cloud accordingly.

The term filter functionality can be used to focus entirely on the co-occurrences by removing all terms from the word cloud that do not share any sentence with the selected ones. Thus, an effective drill down to relevant information is supported that facilitates iterative analytic processes [21]. It is a major feature of our approach and provides a flexible focus-and-context technique for the users.

Another way of adding terms to the term filter is by using the search box ⓒ. Here, users can enter terms independently of whether they are part of the word cloud. If a word is entered that occurs in the text, the term statistics panel ⓓ and the info panel ⓔ display information about it. If it is part of the word cloud, it is highlighted, along with all co-occurring terms. Users can add words from the search box to the term filter list. The search box thus allows the construction of co-occurrence clouds for terms whose frequency is too low to be displayed in the initial word cloud.

3.7 Part-of-Speech and Named-Entities

The filter tab ⓕ enables analysts to explore the word cloud according to different parts-of-speech (POS) and named entities (NE, labeled with 'categories' in the GUI). All POS categories of the Penn Treebank tagset [20] are supported by the CoreNLP POS Tagger [33]. They are condensed to nine major POS categories in the user interface of the Word Cloud Explorer. Likewise, the NE categories of the CoreNLP NE Recognizer [7] are listed in the filter interface.

Users can hover over the POS and NE categories to highlight all respective terms in the cloud and see which terms are annotated which way. As for the co-occurrence highlighting (see above), we use a yellow box whose

Figure 4: The initial word cloud with alphabetical layout for one week of Reuters sports news.

saturation indicates what fraction of occurrences of the term have been tagged with the respective category. The two numbers after each category name denote the number of terms of the respective category being part of the cloud and of the overall text corpus, respectively. Disabling one POS or NE category in the filter tab has the effect that counts of this category are subtracted from the overall frequency of the respective terms.

Being able to filter by part-of-speech and named entities is a powerful feature of the Word Cloud Explorer that enables analysts to explore specific aspects of a text. The word cloud can, for instance, be set up to show only locations that occur together with a certain person in the text, or analysts can look at adjectives that co-occur with a specific organization.

3.8 Text Viewer and Stopword Editor

The text viewer ⓗ lists all sentences in the text that contain the selected and/or focused terms (as shown for the term “looked” in Figure 1). This allows analysts to review terms in their original context and verify hypotheses about the content of the text, e.g. about the connection between two persons.

The text viewer has a very limited functionality in the current implementation. It might be extended to show more text context and allow for sophisticated searching, filtering, and further interaction on its own. Another option would be to reuse an existing text viewer and integrate it into the system.

With the stopword editor ⓖ, users can modify the stopword list and adapt it to different analysis contexts. Any modification of the list triggers an instant update of all views. The list can also be externally changed by editing the corresponding text file.

3.9 Adjustment of the Word Cloud

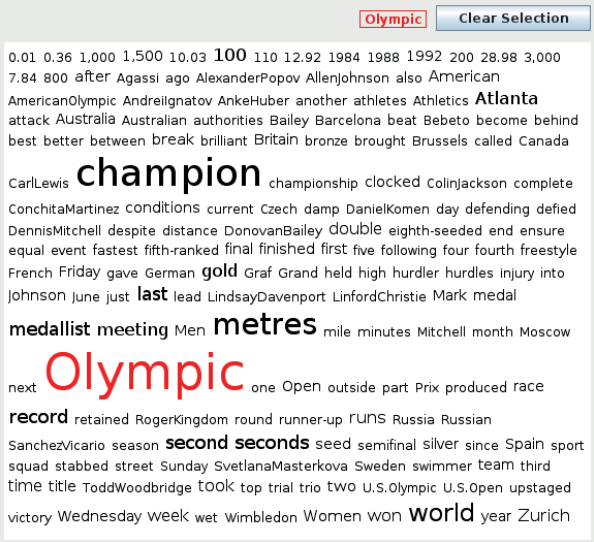
The cloud control panel ⓘ allows to dynamically change the size of the word cloud and of the displayed terms. Users can set the maximum number of terms, the maximum font size of terms, or the minimum frequency that terms require to appear in the word cloud. When one of these values is changed, the other values are adapted accordingly.

In the menu of the system, users can disable stopword filtering, define a cutoff frequency for the co-occurrence calculations, disable the concatenation of multiword expressions, and turn the lemmatizer off. The latter has the effect that different word forms are no longer merged but that each term represents a single word form, as in common word clouds.

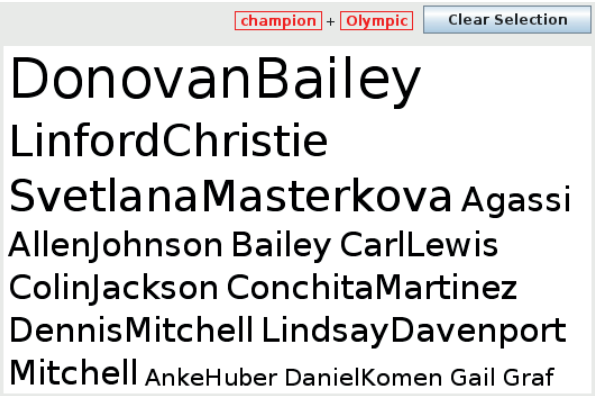
Another feature offered in the menu is a coloring of the terms according to their most frequent part-of-speech. A screenshot of a word cloud with this functionality activated is depicted in Figure 3.

4 Application Example

In the following, we will present an application example that demonstrates the analytical power of the Word Cloud Explorer. The example corpus is Reuters Corpus Volume 1 (RCV1) [17]. It consists of a large collection of manually categorized news articles made available by Reuters Ltd. for research purposes. We use the first week of articles from the corpus, ranging from August 20 to August 26 of the year 1996.



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



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

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

Figure 5: Word clouds of the application example.

Three major global sports events were dominating the news during that week. First, the summer Olympic games that took place in Atlanta, GA that year. Second, the Wimbledon Tennis Championships taking place in London a little earlier. Third, the U.S. Open that started in New York in that week. We therefore restrict the Reuters corpus to sports news by selecting all articles that have been categorized accordingly. As analysis example, we use a task that was also part of the qualitative user study reported in the next section: "Who are the most frequently mentioned male and female Olympic champions during that week?"

In the Word Cloud Explorer, we first choose an alphabetically ordered word cloud as shown in Figure 4. The most frequent terms in this cloud are common words from the domain of sports, such as *won*, *played*, *match*, and *game*. We can also spot some names of countries, cities, and sports events, such as the aforementioned *Wimbledon*, *Olympic* games, and *U.S. Open*. As we are interested in Olympic champions, we hover over the term *Olympic*. In the term statistics panel, we see that it occurs in 87 sentences of the text corpus. The info panel further informs us that it occurs 90 times in those 87 sentences, seven of which are occurrences within multiwords like *Olympic Committee*.

We switch to the co-occurrence cloud of *Olympic* (Figure 5a) and see that the term *champion* is used most often with it. Among the less other terms in the cloud, there are many athletes with their first and last names contracted by the multiword feature. We further see numbers denoting scores, years, distances, etc.

Next, we add *champion* to the term filter and choose the frequency based ordering. We then use the named entity filter to show only persons in the cloud (Figure 5b). We can see right away that *Donovan Bailey* is the most frequently mentioned Olympic champion in the analyzed part of the corpus. To read more about him, we open the text viewer that lists all sentences containing his name. We learn that he is a Canadian sprinter who set a speed record at the 1996 Olympic games.

Looking at the other names in the frequency ordered co-occurrence cloud of Figure 5b, we can quickly spot the female Olympic champion most often mentioned. Her name is the third term in the cloud: *Svetlana Masterkova*. We select this term and deselect the *Person* filter to get the co-occurrence cloud shown in Figure 5c. This cloud reveals further information about Svetlana Masterkova, for instance, that she seems to be Russian and that there appears to be some connection to the distances of 800, 1,000, and 1,500 meters.

To verify these assumptions and get further information, we switch to the text viewer and learn that Svetlana Masterkova is indeed a Russian sprinter and dou-

ble Olympic champion in the 800 and 1,500 meter distances. We also read that she set two world records shortly after her double win at the Olympic games, one in the 1,000 meter track at a competition in Brussels and another in the mile distance of the track and field event Weltklasse Zürich. Those four wins within a short time span caused a lot of attention in the news during the week in focus.

The application example showcases some of the strengths of word cloud based text analytics. At the beginning of the analysis, users quickly get a rough idea of a text's term space by skimming through the word cloud. The most frequent terms and topics are conveyed immediately, offering a good starting point for the analysis. We started the analysis with the term *Olympic* in the application example.

The different types of word clouds facilitate the structured exploration by letting users arrange the term space according to their information needs. In the example, the frequency-based layout helped to identify the persons most often mentioned together with the terms *Olympic* and *champion*. Using the co-occurrence clouds, users can filter the term space and analyze parts of it in more detail, as we did with the name *Svetlana Masterkova*. Co-occurrence highlighting, on the other hand, was useful to interactively explore and discover term relations. The named entity feature allowed us to focus only on person names. Finally, we could use the text viewer to refer back to the relevant parts of the original text in order to verify assumptions and get further useful information.

5 Qualitative User Study

We conducted a qualitative user study to gain further insight into the effectiveness of our text analytics approach and its implementation in the Word Cloud Explorer.

5.1 Material and Tasks

We used three corpora in the study which we have already introduced in the previous sections: The first was the novel "The Hound of the Baskervilles" (see Section 3) that served as a training corpus in the study. The second was the Reuters Corpus Volume 1 (RCV1), restricted to one week of sports news, as in Section 4. The third contained the abstracts of all publications of the IEEE VIS conference series from the years 1998–2011 [12]. We selected the latter corpus to include some research-oriented tasks in the study. A part of it was already shown in the co-occurrence cloud of Figure 2. For the Reuters and VIS corpus, we designed questionnaires, each of which contained twelve tasks of varying difficulty. Basically, the tasks can be categorized into three major groups: (1) frequency based tasks, (2) exploration tasks, and (3) tasks asking for specific terms, with many of the tasks belonging to more than one of these groups. Example tasks include "Who was mentioned most often in connection with 'Ferrari'?", and "What is the Nyquist theory about?".

5.2 Procedure

The participants were five members of the university's visualization institute. They were between 25 and 31 years of age and mostly male (one was female) with very good to excellent English skills according to their own judgment. All participants had experience with expert systems and analysis tasks, and most were familiar with the topic of text analytics. We considered this beneficial as we aimed for informed feedback from experts in this area.

The procedure for each of the participants consisted of the following five steps: *i) color vision deficiency test:* Each participant was tested for color vision deficiencies with the Ishihara color plates. *ii) user training:* We explained the features of the Word Cloud Explorer using the "The Hound of the Baskervilles" corpus. The participants could ask questions and try out the system until they felt confident to use it. *iii) task completion:* We asked the participants to solve the analysis tasks, beginning with the Reuters sports news corpus followed by the corpus with the VIS abstracts. In case participants were stuck during a task, we kept hints on a separate sheet of paper that they could consult. To get additional insight, we encouraged the participants to articulate their thoughts during task completion according to the think-aloud method. *iv) questionnaire:* We asked the participants to complete a questionnaire with demographic questions and questions about their experience with and thoughts on the approach. *v) discussion:* Finally, we discussed the general approach and specific analysis tasks with the participants.

5.3 Results

The study participants were surprised by the possibilities offered by such a simple and straightforward visualization as word clouds if enriched with context information and sophisticated interaction techniques. Overall, they performed very well in solving the analysis tasks with the Word Cloud Explorer. The paper with the hints was hardly ever used and the tasks were generally solved quickly and correctly without any help.

The participants rated the Word Cloud Explorer as an intuitive and useful text analysis system. They stated that they could imagine using it to analyze large bodies

of text. However, they also stressed that they would only use it in combination with other tools complementing its functionality. For instance, while the ability to refer back to the actual text source was considered a crucial feature, some participants criticized the limited functionality of the text viewer component and proposed to integrate the approach with a powerful text editor.

With respect to the word cloud layouts, an interesting finding was that all participants preferred the sequential layouts over the circular one, although they rated the circular one to be aesthetically most appealing. When asked about this apparent contradiction, most participants answered that they found it easier to visually compare relative word sizes using the line-by-line layout. This is because the lines could be used as visual anchors which facilitate to compare font height. Furthermore, we could observe the participants switching between the frequency and alphabetically ordered layouts according to whether they were interested in high frequency terms or searching for a specific term. This indicates that it is important to provide different word cloud layouts that users can choose from depending on the analysis task.

However, most participants preferred the search box over the word cloud when searching for a specific term. This result is not surprising, as using the search box is fast and also allows to find terms that are not part of the displayed word cloud. Furthermore, it is in line with the finding of Sinclair and Cardew-Hall [26] that a search box is preferred for specific tasks while a word cloud for more general ones.

The participants were in disagreement about the usefulness of the part-of-speech coloring function. Some considered it a useful feature, while others found that it has little analytical value. It was argued that the part-of-speech of most words does not need to be visually communicated, as users normally know the part-of-speech of a word once they read it. Using the part-of-speech categories as a filter for the word cloud was considered more useful.

The named entity feature was unanimously found helpful and the participants used it frequently to solve the analysis tasks. The aggregation of multiwords and different word forms was also positively evaluated by the participants. Overall, the participants assigned the linguistically and interactively improved word clouds many positive attributes (such as “tidy”, “clear”, “efficient”, “useful”). All this feedback indicates that word cloud based approaches have indeed some potential in text analytics if part of an advanced implementation like the Word Cloud Explorer. The user study revealed that their main advantages are flexibility and intuitiveness, which indicates that they might be particularly beneficial in environments were training times should be minimal.

6 Conclusion and Future Work

In this work, we explored the extension of the basic word cloud visualization with additional information and interactive features to transform it into a powerful tool for text analytics. As proof of concept, we developed the Word Cloud Explorer, a prototypical system that uses word clouds as its central visualization method and integrates several interactive features into one consistent framework for interactive text analysis. We demonstrated the applicability of the approach in an example and evaluated it in a qualitative user study. The study results indicate that word clouds are indeed an effective tool for text analysis if equipped with further information and sophisticated interaction techniques.

In future work, we plan to address the handling and comparison of multiple documents. An interesting question in this respect is how to extend the word cloud view to allow for the comparison of several documents at a time. This could, for instance, be done using different colors, similar to the word cloud component of the ManyEyes website [35]. Other alternatives could be Parallel Tag Clouds [4] or the use of a word cloud matrix [36, 39]. In general, we aim to integrate the presented approach with related work to allow for even more comprehensive text analysis. We are aware that this has to be done carefully, because, as this work indicates, simplicity and intuitiveness are important strengths of word cloud based text analytics.

Acknowledgements

This work was partially supported by the German Science Foundation (DFG) in the context of the priority program ‘Scalable Visual Analytics’ (SPP 1335).

References

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

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

[3] Y.-X. Chen, R. Santamaría, A. Butz, and R. Therón. TagClusters: Semantic aggregation of collaborative tags beyond tagclouds. In *Proc. 10th Int. Symp. Smart Graphics*, SG ’09, pages 56–67. Springer, 2009.

[4] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. of IEEE Symp. Visual Analytics Science and Technology*, VAST ’09, pages 91–98. IEEE, 2009.

[5] 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.

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

[7] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*, ACL '05, pages 363–370. ACL, 2005.

[8] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda. Topigraphy: visualization for large-scale tag clouds. In *Proc. 17th Int. Conf. World Wide Web*, WWW '08, pages 1087–1088. ACM, 2008.

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

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

[11] Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *Proc. Int. Conf. Multidisciplinary Information Sciences and Technologies*, InScit2006, 2006.

[12] F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE Trans. Vis. Comput. Graphics*, 18(12):2839–2848, 2012.

[13] 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.

[14] 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.

[15] B. Y.-L. Kuo, T. Hentrich, B. M. . Good, and M. D. Wilkinson. Tag clouds for summarizing web search results. In *Proc. 16th Int. Conf. World Wide Web*, WWW '07, pages 1203–1204. ACM, 2007.

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

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

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

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

[20] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993.

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

[22] 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 *Proc. SIGCHI Conf. Human Factors in Computing Systems*, CHI '07, pages 995–998. ACM, 2007.

[23] I. A. Sag, T. Baldwin, F. Bond, A. A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for nlp. In *Proc. 3rd Int. Conf. Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15. Springer, 2002.

[24] J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, CHI '09, pages 2037–2040. ACM, 2009.

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

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

[27] G. Smith. *Tagging: People-Powered Metadata for the Social Web*. New Riders, 2008.

[28] Stanford CoreNLP. A Suite of Core NLP Tools. http://nlp.stanford.edu/software/corenlp.shtml.

[29] J. Stasko, C. Görg, and Z. Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Inf. Vis.*, 7(2):118–132, 2008.

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

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

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

[33] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. 2003 Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology, Volume 1*, NAACL '03, pages 173–180. ACL, 2003.

[34] F. B. Viégas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

[35] 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. Graphics*, 13(6):1121–1128, 2007.

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

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

[38] WP-Cumulus. Word press plugin. http://wordpress.org/extend/plugins/wp-cumulus/.

[39] 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. Graphics*, 16(6):1109–1118, 2010.