DOI: 10.1111/j.1467-8659.2012.03107.x
Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

Semantic Wordification of Document Collections

Fernando V. Paulovich[1], Franklina M. B. Toledo[1], Guilherme P. Telles[2], Rosane Minghim[1] and Luis Gustavo Nonato[1]

[1]ICMC/USP, São Carlos/SP, Brazil
[2]IC/UNICAMP, Campinas/SP, Brazil

Abstract

Word clouds have become one of the most widely accepted visual resources for document analysis and visualization, motivating the development of several methods for building layouts of keywords extracted from textual data. Existing methods are effective to demonstrate content, but are not capable of preserving semantic relationships among keywords while still linking the word cloud to the underlying document groups that generated them. Such representation is highly desirable for exploratory analysis of document collections. In this paper we present a novel approach to build document clouds, named ProjCloud that aim at solving both semantical layouts and linking with document sets. ProjCloud generates a semantically consistent layout from a set of documents. Through a multidimensional projection, it is possible to visualize the neighborhood relationship between highly related documents and their corresponding word clouds simultaneously. Additionally, we propose a new algorithm for building word clouds inside polygons, which employs spectral sorting to maintain the semantic relationship among words. The effectiveness and flexibility of our methodology is confirmed when comparisons are made to existing methods. The technique automatically constructs projection based layouts the user may choose to examine in the form of the point clouds or corresponding word clouds, allowing a high degree of control over the exploratory process.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques— H.5.0 [Information Interfaces and Presentation]: General—

---

1. Introduction

Word Clouds have emerged as a fundamental tool for visualizing document collections. Recent approaches have improved considerably on known issues of the technique, such as the lack of semantic relationship among the words and the inefficient use of space. Word clouds have also been combined with other visual resources such as spark lines and bubble sets for increased effectiveness in practical visual text analysis (VTA) applications.

Despite this progress and the clear success of the technique given its widespread use, there is potential for improvements aimed at document analysis applications. For example, existing methods do not yet provide an intuitive visual representation that allows to link words on the layout to the documents they are meant to represent. In other words, the simultaneous visualization of the neighborhood relationship among documents and their corresponding words in the cloud is not offered by existing methods. A mechanism for relating groups of documents to corresponding word clouds in a flexible and interactive manner will allow a familiar and proven visual representation to offer support to a larger number of high-level tasks and applications of document analysis.

Another aspect that has not been properly addressed by previous methods is the construction of word clouds inside general polygons with semantical preservation between words. More precisely, some methods can build semantically consistent word clouds, but not inside general polygons. On the other hand, techniques capable of fitting word clouds inside polygons do not preserve the semantic relationship among keywords. Combining these features together will empower the word cloud visual representation considerably. With this capability, various semantically consistent word clouds can be visualized simultaneously, by just dividing the visual space into a polygon tiling, allowing examination and correlation of many document collections at once. Additionally, polygon based displays of word clouds are necessary to have a direct correspondence between word clouds and general 2D layouts of document collections produced by point placement strategies, such as MDS (Multidimensional Scaling) or multidimensional projection plots, with groups of interest bounded by polygons.

This work proposes a novel word cloud-based visualization technique, named ProjCloud, which presents solutions to the two issues raised above. The matter of visualizing word clouds and their corresponding documents simultaneously is tackled by combining a multidimensional projection with a new algorithm for positioning keywords inside polygons. Such combination is achieved by building word clouds based on the coordinates resulting from a multidimensional projection. Similar documents are identified in the visual space and polygons enclosing their groups are formed, either automatically or manually, to define regions in which the wordification takes place. That gives an immediate association between positions of keywords and their underlying documents.

The word placement scheme we propose is based on an approximation to the solution of the *cutting-stock optimization problem*, and generates pleasant arrangements of words with efficient use of space. The semantic relationship among keywords is established using a graph-based spectral sorting scheme that defines the order in which words are positioned inside a polygon. Besides defining the semantic relation, the spectral sorting scheme also provides a reliable mechanism for adding weights to the relevance of words. The resulting relevance identification scheme is very effective to highlight the important keywords, making the word cloud more informative and easier to analyze.

In summary, the main contributions of this work are:

- A novel methodology for combining multidimensional projection and word clouds, which enables to visualize the similarity among documents as well as their corresponding word clouds in an integrated manner, extending the exploratory capabilities of word clouds.
- A new approach for building word clouds inside polygons while still preserving the semantic relationship among keywords.
- A mechanism based on spectral sorting that allows arranging words according to their semantic relationship as well as highlighting the most important words in the cloud.

The contributions above endow the proposed techniques with a set of traits that are not present in any other word cloud-based visualization technique. To the best of our knowledge, this is the first methodology that provides a mechanism enabling semantic distribution of word clouds in general polygons while combining this semantic with relevance information and with indexing of underlying documents.

2. Related Work

The literature on visualization of document collections is quite extensive and currently available methods vary greatly in regard to the mathematical and computational frameworks employed to assist textual document analysis. In order to contextualize the technique proposed in this paper we focus our discussion only on techniques that rely on word or tag cloud paradigm to perform text visualization.

One of the first approaches to use keywords as a visualization resource was proposed by Kuo et al. [KHGW07]. It builds a word cloud inside a rectangular box using a line-by-line arrangement combined with a scaled-by-relevance scheme to highlight the importance of each word. The line-by-line scheme is prone to produce a large amount of white space, a problem tackled by Kaser and Lemire [KL07] through a packing mechanism. The white space problem was drastically reduced by the spiral-based arrangement scheme called Wordle [VWF09], which not only scales words according to relevance but also rotates them to make a better use of the available space. Despite the pleasant layout and the efficient use of the white space, Wordle does not take into account any semantic relationship among words when positioning them in the layout.

Although the brain is capable of grasping meaning from a set of seemingly unrelated words, the maintenance of semantic relationships can support more complex analyzes, reduce ambiguity of meaning and speed up users' actions towards examining the content of a group of documents from their word cloud. To resolve the problem of semantic association among words, some techniques have employed alternative arrangements such as circular layouts [SB07], which enable to incorporate some semantic relationship between words but at the price of generating additional blank space between them. User interaction was the solution proposed by ManiWordle (Manipulable Wordle) [KLKS10] to add semantic into Wordle clouds. This manipulation enables flexible control of the layout while ensuring a pleasant word arrangement. The technique proposed by Cui et al. [CWL*10] combines a trend chart and a force directed scheme in order to keep the semantic relationship among words while visualizing the temporal structure of documents.

Other methods such as SparkClouds [LRKC10], Parallel Tag Clouds [CVW09], and Document Cards [SOR*09] aim at augmenting word clouds with extra visual resources such as spark lines, parallel coordinates, and images, respectively. Their goal is to improve word clouds in their ability to convey information.

The issue of word positioning within polygons was addressed by Seifert et al. [SKK*08] through an divide-and-conquer heuristic mechanism and by Shi et al. [SWL*10], who use a line-based scheme that starts the scan process from the centroid of the polygon. Their technique, though, cannot ensure semantic arrangement of words inside the polygon. More recently, Wu et al. [WPW*11] presented a methodology that first computes the semantic relationships and then uses multidimensional scaling to place those words on visual space, removing blank spaces through a carving scheme. However, their technique cannot enforce words to lie inside geometrical shapes. Additionally, the correspondence between words and documents is only possible through a spreadsheet-like visualization mechanism, which does not enable a panoramic view of how documents relate to each other and is unconnected to the layout of the word clouds.

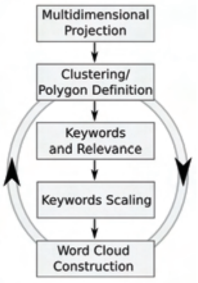
The technique presented in this paper encompasses a set of traits that cannot be found simultaneously in any other word cloud-based document visualization technique and are meant to surpass some of their limitations.

Firstly, semantic is preserved in local sense, by word ordering, and in a global sense, that is, the user has a well established correspondence between words and documents. Another important aspect is that the user can interactively select and examine a subset of documents and their neighborhood while associating them to the word cloud layout. The arrangement is done inside general polygons created from a projection layout. The definition of the polygons can be done automatically or interactively, adding exploration flexibility to the approach.

3. Wordification Technique

Before going into the technical aspects of our approach we provide an overview of the sequence of steps that accomplishes our wordification technique.

ProjCloud starts by mapping the document collection into the visual space using a multidimensional projection technique. In our implementation we employ the *Least Square Projection (LSP)* [PNML08] to perform such mapping. LSP takes a vector space representation of the text set as input [Sal91]. Any other multidimensional projection techniques that is capable of handling large data sets with low computational cost can be employed as well (see [JCC*11] for an up-to-date survey of efficient multidimensional projection methods). In the second stage of the pipeline, the groups of documents to be wordified are defined. In the automated version of this step, points in the visual space are clustered using the bisecting k-means algorithm [SKK00] and the convex hull of each cluster is computed, thus segmenting the data into a set of non-overlapping convex regions. If the system is running in interactive mode then the user can freely draw polygons containing the groups of data points to be wordified. As illustrated in the inset on the right, in the third stage of the pipeline, keywords are determined from the most frequent words related to the documents contained in each polygon, and their relevance computed in order to guide the semantic-preserving placement of words (Subsection 3.1). Once relevance has been established, the scaling step takes place, that is, keywords are sized based on their relevance and on the area of the containing polygon (Subsection 3.2). In the final step of the pipeline, the keywords are placed inside the polygons using the optimization scheme described in Subsection 3.3. The optimization algorithm takes into account the semantic relationship among keywords to produce a semantic-preserving layout.



3.1. Keyword Relevance and Semantic Relation

Let M be the document x term frequency matrix (see [Sal91] for details on how to compute M) containing the n most frequent terms or keywords in a set of documents contained in a polygon P ($n = 200$ in our implementation and it represents the maximum number of keywords in a cloud). Denoting the subset of keywords by $\Phi = \{\phi_1, \ldots, \phi_n\}$ and by C the covariance matrix obtained from M, we build a graph G where each node corresponds to a keyword ϕ_i and an edge e_{ij} connects a node ϕ_i to the node ϕ_j if only if the covariance c_{ij} (i, j entry in C) is among the k-largest ones in the row i or row j of C ($k = 10$ in our implementation). Assuming that edge e_{ij} has weight c_{ij}, it is well known from the literature that the second eigenvector of the weighted graph Laplacian derived from G [Chu97], called Fiedler vector, assigns a scalar value α_i to each node ϕ_i that minimizes:

$$\min \sum_{e_{ij}} c_{ij}(\alpha_i - \alpha_j)^2 \tag{1}$$

In other words, $(\alpha_i - \alpha_j)$ will be small when c_{ij} is large, that is, nodes ϕ_i and ϕ_j will receive similar values when they are closely related. Therefore, if the keywords are sorted according to α_i, semantically correlated ones will be placed close to each other in the sorted sequence.

The most relevant keyword from the subset Φ is defined as follows. Let c_{ij}^{max} be largest entry in C and ϕ_i^{max} and ϕ_j^{max} be the corresponding keywords, that is, ϕ_i^{max} and ϕ_j^{max} are the keywords with larger covariance. The most relevant keyword is ϕ_i^{max} if the average covariance between ϕ_i^{max} and ϕ_k is larger than the average covariance between ϕ_j^{max} and $\phi_k, k = 1, \ldots, n, k \neq i, j$, otherwise the most relevant keyword is ϕ_j^{max}. If the averages are equal, either ϕ_i^{max} or ϕ_j^{max} can be chosen as the most relevant keyword.

Once the most relevant keyword has been defined, say ϕ_r, the keywords are sorted in increasing order according to $||\alpha_r| - |\alpha_k||, k = 1, \ldots, n, k \neq r$, which sets the relevance of all keywords. The Fiedler-based approach also allows the removal of words with low relevance. If $||\alpha_r| - |\alpha_k||$ is larger than a threshold then ϕ_k is not included in the final list of relevant words. In our experiments, a threshold equal to $1.2 * ||\alpha_i^{max}| - |\alpha_j^{max}||$, where α_i^{max} and α_j^{max} are the Fiedler values corresponding to ϕ_i^{max} and ϕ_j^{max} respectively, was enough to produced the results presented in next section. In ProjCloud the order given by the Fiedler vector dictates the position of the words into the cloud. The size of the words also depends on the Fiedler vector, as detailed below.

3.2. Sizing Keywords

The size of each keyword is defined as follows. First the scalar values $c_{rk}/(1+||\alpha_r|-|\alpha_k||)$ are scaled to fit in the interval $[f_{min}, f_{max}]$, where f_{min} and f_{max} are the minimum and maximum sizes of the text fonts respectively. Values of $f_{min}=12$ and $f_{max}=50$ prove to form a good range in our experiments. The font size of each keyword ϕ_k is then set to the scaled value of $c_{rk}/(1+||\alpha_r|-|\alpha_k||)$. If the summation of areas of all keywords' bounding boxes is smaller than the area of the polygon P, f_{max} is increased and the values are re-scaled. This process is repeated until the sum of areas of the keywords exceeds the area of P.

3.3. The Optimization Problem

Mathematical Formulation. Given the set of keywords $\Phi=\{\alpha_1,\ldots,\alpha_n\}$ scaled with respect to a polygon P, the problem of constructing a word cloud from Φ constrained by P consists in placing each keyword ϕ_i inside P in such a way that it avoids overlaps and keeps the space between words as tight as possible. Such a problem becomes more manageable if we replace the keywords by their corresponding bounding boxes. In fact, the problem of filling a polygon P with rectangular boxes of distinct sizes is a variant of the two-dimensional version of the *Cutting Stock Problem* (CSP) [WHS07]. The CSP is a classic combinatorial optimization problem in which rectangular parts of various lengths, possibly rotated by 90^o from one another, must be cut from a plate with minimum trim-loss. This problem has been shown to be NP-Complete [FPT81], thus exact solutions in reasonable time can only be found for problems involving a small number of elements.

Heuristic algorithms have been the alternative adopted to approximate large CSP problems [HT01]. In this work we propose a variant of the heuristic proposed by Baker et al. [BCR80], which is efficient enough to enable interactive manipulation of word clouds and flexible enough to allow incorporating the semantic relationship among words during the word cloud construction process.

The proposed algorithm for fitting rectangular boxes into a given polygon can be stated as follows:

The Optimization Algorithm. Assume that the list of rectangular boxes $\Phi=\{\phi_1,\phi_2,\ldots,\phi_k\}$ is ordered according to their relevance (from now on we do not make any distinction between the keywords and their corresponding bounding boxes). The elements ϕ_k are positioned inside the polygon P in the given order and positioned as close as possible to the centroid of P, avoiding overlap. Starting with ϕ_1, the scaling mechanism described in Section 3.2 provides an initial value for the size of ϕ_1. If ϕ_1 does not fit horizontally within P then it is rotated 90^o and a new attempt is made. If the rotated element also does not fit in P then ϕ_1 is scaled down by a factor δ and the whole process is repeated until ϕ_1 fits inside P. In our implementation the value of δ was set to 0.925.

Suppose now that the elements $\phi_1,\ldots,\phi_{j-1}$ have already been placed inside P. In order to position the next element ϕ_j we first define a background regular grid covering P with cell size equal to Δx. The value of Δx dictates the minimum amount of blank space between words. Cells in the background grid covered by an element ϕ_l, $1 \leq l \leq j-1$ are marked as not-available. If there is a subset of available cells in the background grid where ϕ_j, or a rotated version of ϕ_j, can lay on, then ϕ_j is placed. If there is not enough room in P, then $\phi_1,\ldots,\phi_{j-1},\ldots,\phi_k$ and the background grid are scaled down by δ, the cells of the new grid are labeled as either available or not-available, and the whole process is repeated. The search for a place for ϕ_j starts from the available grid cells that are closest to the centroid of the polygon. This priority mechanism combined with the Fiedler's order tends to keep semantically related words next to each other in the final layout.

It is important to point out that the background grid avoids expensive intersection calculations, rendering the algorithm quite efficient. The pseudo-code shown in Algorithm 1 summarizes the main steps of the algorithm.

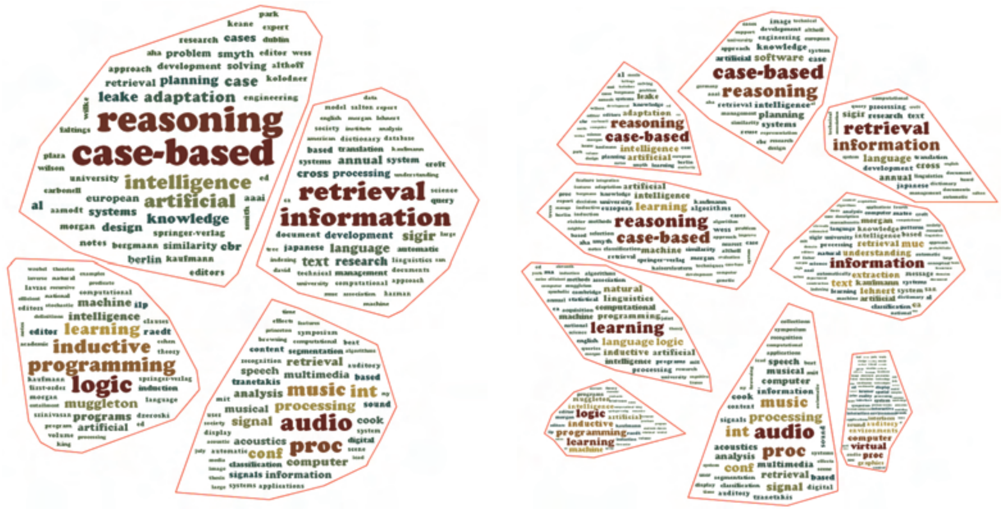
4. Results and Comparisons

As mentioned in Section 3 our approach allows for automatic as well as user-driven construction of word clouds. In this section we show that both approaches can be quite useful when analyzing and exploring document collections.

Assuming as input a set of documents and their projection on the visual space, the automatic version of our method starts by clustering the documents in the visual space using the bisecting k-means algorithm. Polygons surrounding each cluster are defined as the convex hull of the cluster. A word cloud is then built inside each convex polygon using the algorithm described in Subsection 3.3.

Figure 1(a) shows the result of applying this automatic wordification mechanism on a collection of 675 scientific articles in four distinct areas: case-based reasoning, inductive logic programming, information retrieval, and sonification. It can be seen that the distinct groups of documents are easily identified and the main topics describing the content of these documents are clearly highlighted.

An interesting aspect of ProjCloud is that the bisecting k-means splits the clusters recursively, thus word clouds can naturally be refined to show more detailed information inside each cluster, as illustrated in Figure 1(b). Notice that when we choose to define nine clusters, ProjCloud automatically splits the top left cloud in Figure 1(a) in three new clouds,



(a) Four Groups.

(b) Nine Groups.

Figure 1: *Visualization of a document collection generated automatically by ProjCloud from a collection of scientific papers in four different areas of knowledge.*

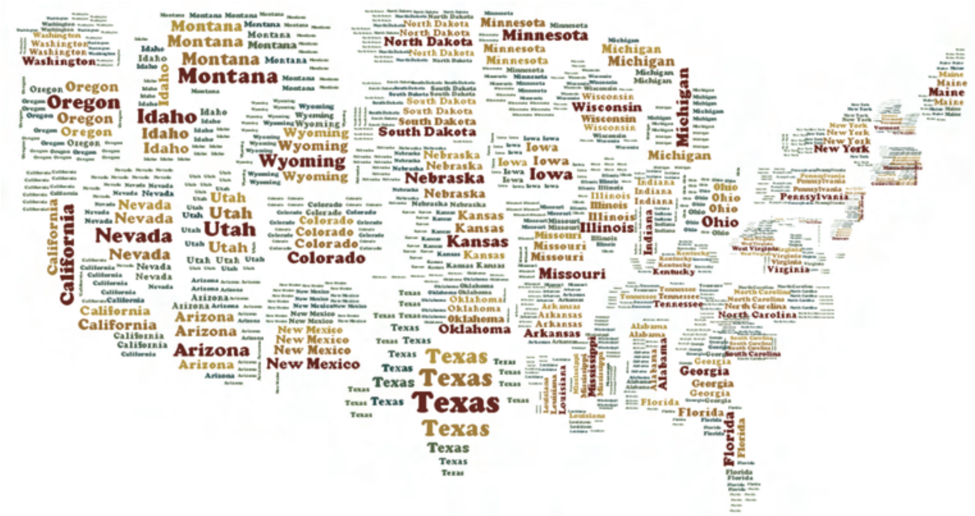
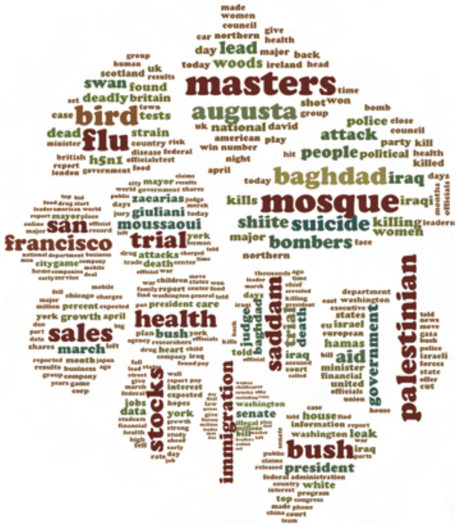


Figure 2: *ProjCloud behavior for complex polygon shapes.*

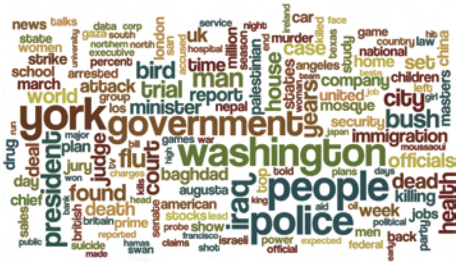
namely the three top left clouds in Figure 1(b), while the three other clouds are each split in two.

Figure 2 shows that even non-convex complex polygonal shapes can satisfactorily be handled by ProjCloud. This example has been generated by taking the polygons that define the boundaries of U.S. states and arranging 25 words with random weights inside each polygon.

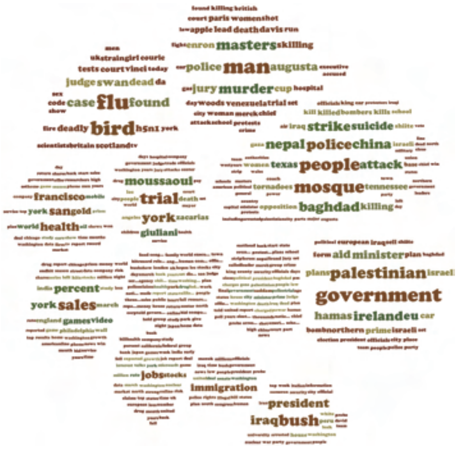
The importance of clustering the data before building the word clouds becomes evident in Figure 3, where word clouds generated by ProjCloud and Wordle [VWF09] are presented. Those clouds have been produced from a data set containing 2,625 RSS news feeds from CNN, BBC, Reuters and Associated Press collected during two days in April 2006. Notice from Figure 3(b) that although Wordle produces a pleasant layout, it is harder to realize what the actual contents of the documents are. In contrast, ProjCloud displays sets of documents according to their topics and builds semantically consistent word clouds in each group of documents, making it easier to conceptualize a mental model of



(a) ProjCloud of the entire news collection.



(b) Wordle of the entire news collection.



(c) Word cloud using the approach proposed in [SKK*08].

Figure 3: *Comparison between ProjCloud, Wordle, and the approach proposed by Seifert et al. [SKK*08] when visualizing the RSS news feed dataset. The semantic information conveyed by ProjCloud, which is not preserved by Wordle and the Seifert's approaches, makes it easy to interpret the content of this document collection.*

Algorithm 1 Polygon coverage algorithm.

```
input:   - $\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$: the list of ordered keywords.
         - $P$: the polygon to place the keywords.
         - $c_P$ : the polygon centroid.
output:  - the placement of $\Phi$ inside $P$.

procedure PolygonCoverage($\Phi, P, c_P$)
 1: set the background grid with cell size $\Delta x$
 2: while $\Phi \neq \emptyset$ do
 3:     $\phi \leftarrow$ the first term in $\Phi$
 4:     Remove $\phi$ from $\Phi$
 5:     Find a subset of available cells $c$ to place $\phi$
 6:     if $c$ was found then
 7:         Place $\phi$ on $c$ and mark it as not-available
 8:     else
 9:         Rotate $\phi$ by $90^0$
10:         Find a subset of available cells $c$ to place $\phi$
11:         if $c$ was found then
12:             Place $\phi$ on $c$ and mark it as not-available
13:         else
14:             Scale down $\Phi$ and the background grid by $\delta$
15:             Recreate the grid
16:             Re-label the not-available cells and go to step 5
17:         end if
18:     end if
19: end while
```

the data from the visualization. For instance, in Figure 3(a), it is easy to perceive that the most important news are clearly visible, such as an event related to bird flu, the trial and sentence to death of Saddam Hussein, as well as other news.

In Figure 3 we also compared our approach with the one propose by Seifert et al. [SKK*08], which, similarly to the strategy proposed here, seeks to place words inside enclosing polygons. In their work, the authors define three operations, namely shifting, font size scaling and word text truncation. With these, they managed to define a family of algorithms for word cloud construction. Figure 3(c) presents the resulting word clouds for the RSS news feeds dataset, considering the groups of documents and the polygons conveyed by our technique, with frequency counting as the strategy to define the terms relevances. These were built using the *trunc-scale-shift* algorithm with initial font size and font thresholds as suggested in the original work. We also adopted the convention that, if the algorithm stopped without placing all words defined for a cloud then we use the placement with the maximum number of words as the answer. In Figure 3(c) we can see that clouds with narrow shapes or many words are the mostly visually hampered. Both truncation and font size flattening contributed to that. In particular, truncation makes some words become unrecognizable. Adopting a first fit strategy, by which not every word is placed by the algo-

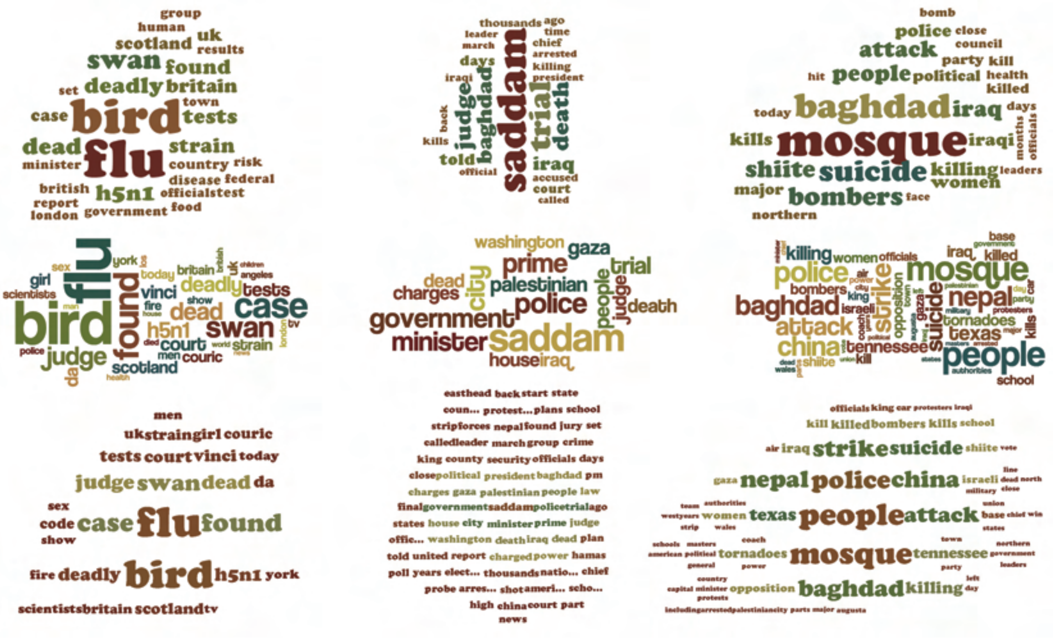


Figure 4: *Comparison between ProjCloud (first line), Wordle (second line) and Seifert et al. [SKK*08] approaches (third line) for layouts when visualizing RSS news feeds from focused groups.*

rithm, results in clouds with a few words only (figure not shown), and reducing the font sizes intervals impairs the visualization as well, while allowing more words to be placed. Specially for narrower regions, the effects of the lack of word rotation are easier to notice.

In our approach, the conveyed semantic is a consequence of the efficient relevance identification and weighting mechanism resulting from the proposed Fiedler-based sorting scheme, that improves the more common frequency based mechanism. In addition, the good layout results from the our adapted fitting algorithm that interactively scales and rotates the words to fill the available area as much as possible, without discarding or trimming them.

Figure 4 shows in more detail the effectiveness of the Fiedler-based sorting scheme in preserving the semantic relationship among words (first line in the picture). The second line shows the word clouds generated by Wordle and the third line shows the word clouds produced by Seifert's approach from the same subset of documents of the RSS news feeds data file shown by ProjCloud layouts in the first line. One can notice that only the cloud containing the words “Bird” and “Flu” conveys meaningful information and it is more difficult to conclude what is the actual subject of the other two clouds. Also, the problems of trimming words and flattening font sizes become evident on the layouts produced by Seifert's approach. On the other hand, the word clouds produced by ProjCloud clearly highlights the main themes of the news due to the semantically consistent placement of words.

The advantage of combining multidimensional projections and word clouds becomes evident when the user is allowed to interact with the collection under analysis. Figure 5 illustrates the possibilities of our system by showing how word clouds can be generated interactively, where the user selects two specific groups of documents on the projection of the documents. In this example the IEEE Infovis 2004 contest data set [FGP04] with 615 papers was used. Since a multidimensional projection allows the visualization of neighborhood relationships among points, users can interactively select groups of documents from which the word cloud is built, a functionality not present in any other visualization system based on word cloud.

5. Discussion and Limitations

The results presented in Section 4 show the effectiveness of ProjCloud as a visualization tool for analyzing and exploring document collections. The capability of refining word clouds by increasing the number of clusters has turned out to be very useful in practice, allowing for overview and detail interaction, and supporting exploration based on topics of groups and subgroups of documents through a tool that is intuitive and easy to use. Moreover, the refinement process preserves the context of the word clouds, since each word cloud in a finer level derives from a single cloud in the pre-

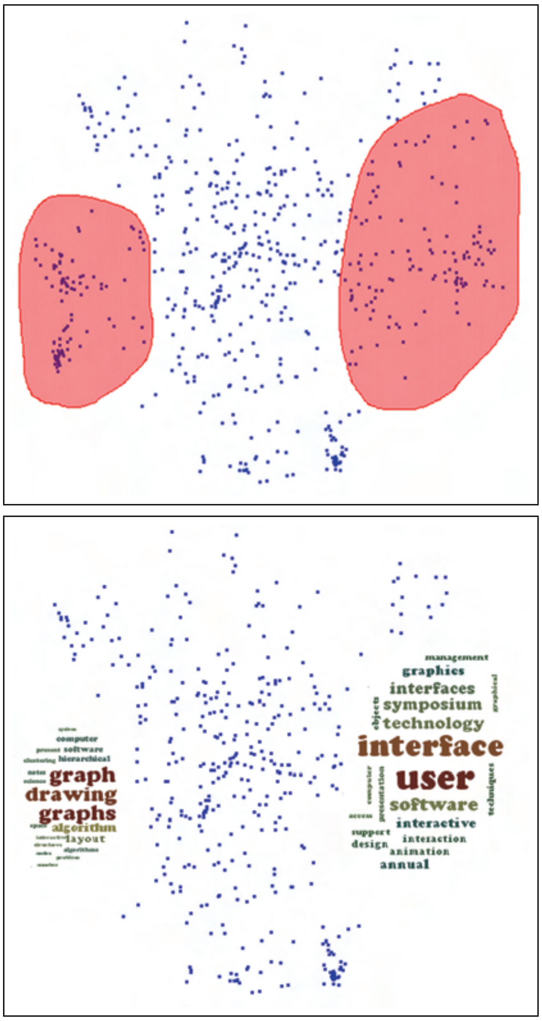


Figure 5: *A user can interactively draw a region (polygon) containing a subset of documents of interest (top figure). Keywords are extracted from the selected document and their corresponding word could is built inside the user-defined region (bottom figure).*

vious level. As far as we know, ProjCloud is the first technique to enable such a multilevel visualization mechanism for word clouds.

The Fiedler-based sorting scheme is another interesting mechanism introduced by ProjCloud. Besides enabling the construction of semantically consistent word clouds, the spectral scheme is also very reliable to define the relevance of words. Notice for example in Figure 1 that fonts vary quite abruptly. At first glance such abrupt change in the size of the words might seem a weakness of ProjCloud. However, what is happening in fact is that the spectral sorting mechanism assigns considerably smaller values to words that are less relevant. These less relevant words become much smaller than the most relevant ones, easing visual identification of the main topics of the underlying documents.

An aspect to be observed is that the size of each polygon derives from the geometric location of its corresponding cluster in visual space, which enforces the word sizing mechanism to be local. More precisely, if the largest word ϕ_{P_1} inside a polygon P_1 is smaller than the largest word ϕ_{P_2} in a polygon P_2 then ProjCloud ensures that ϕ_{P_1} and ϕ_{P_2} are the most relevant words in the clusters that gave rise to P_1 and P_2. However, one cannot claim that the relevance score of ϕ_{P_1} is smaller than the relevance score of ϕ_{P_2}. In other words, ProjCloud does not provide a direct mechanism to compare the relevance of words from cluster to cluster. The information about the relevance of keywords can be conveyed, though, by coloring the boundary edges of the polygons according to the number of instances inside that polygon or to the score of the most relevant keyword.

ProjCloud also offers a new perspective to word cloud based analysis brought by the use of integrated multidimensional projection techniques. In fact, multidimensional projections have long been used to analyze an explore documents and textual data [PNML08]. However, visual resources employed in combination with projections are still restricted to points and their visual attributes and textual tags to help in the identification of individual documents. We believe that the combination of multidimensional projections and word clouds as proposed by ProjCloud opens a line of approaches for visualization and exploration of textual document collections.

ProjCloud is largely dependent on the clustering process. More precisely, if the clustering performs poorly, for instance by causing concentration of points in specific regions of the visual space, then the associated convex polygon will be too small, thus making the word cloud difficult to fit and read. Although zooming can be used to mitigate the problem, dependence of the clustering scheme is still an issue that we plan to address in the near future. Another aspect to be tackled in future versions of the system is the “void” space between clusters. Although the clustering scheme has the good property of evenly distribute data instances among the cluster, the convex hull mechanism tends to leave too much empty space between them. We are currently investigating a post-processing optimization scheme that scales polygons to change their sizes and minimize the space between them.

6. Conclusions

In this work we propose a novel document collection visualization that combines features of multidimensional projections and word clouds in a single visual environment. ProjCloud produces visualizations where documents are grouped according to their similarity and constructs semantically coherent word clouds for each group of documents. The groups and corresponding word clouds can be refined to reveal subtopics and more detailed information about the documents. Users can also identify documents of interest interactively, thus enabling a richer visualization resource in the context of multidimensional projection.

The mechanism for word distribution inside the polygons presented here is very effective to fit and display words by their relevance, maintaining the semantic relationship between words. The relevance finding mechanism also guarantees that important topics are not missed in the display.

The flexibility and effectiveness of the combination of techniques presented here empower the word cloud mechanism, allowing its use for more demanding text analysis applications.

We are currently investigating better mechanisms to lessen ProjCloud's dependence of the clustering scheme, what should enable a truly multiscale technique.

Acknowledgments

The authors acknowledge the financial support of the Brazilian financial agencies CNPq and FAPESP.

References

[BCR80] Baker B. S., Coffman E. G., Rivest R. L.: Orthogonal Packings in Two Dimensions. *SIAM Journal on Computing 9*, 4 (1980), 846–855. doi:10.1137/0209064. 4

[Chu97] Chung F. R. K.: *Spectral Graph Theory*. No. 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997. 3

[CVW09] Collins C., Viegas F. B., Wattenberg M.: Parallel tag clouds to explore and analyze faceted text corpora. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2009), pp. 91 –98. doi:10.1109/VAST.2009.5333443. 2

[CWL*10] Cui W., Wu Y., Liu S., Wei F., Zhou M., Qu H.: Context-preserving, dynamic word cloud visualization. *IEEE Computer Graphics and Applications 30* (2010), 42–53. doi:10.1109/MCG.2010.102. 2

[FGP04] Fekete J.-D., Grinstein G., Plaisant C.: IEEE InfoVis 2004 Contest, the history of InfoVis. www.cs.umd.edu/hcil/iv04contest, 2004. 7

[FPT81] Fowler R. J., Paterson M. S., Tanimoto S. L.: Optimal packing and covering in the plane are NP-complete. *Information Processing Letters 12*, 3 (1981), 133 – 137. doi:10.1016/0020-0190(81)90111-3. 4

[HT01] Hopper E., Turton B. C. H.: A review of the application of meta-heuristic algorithms to 2D strip packing problems. *Artificial Intelligence Review 16*, 4 (2001), 257–300. doi:10.1023/A:1012590107280. 4

[JCC*11] Joia P., Coimbra D., Cuminato J. A., Paulovich F. V., Nonato L. G.: Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics 17* (2011), 2563–2571. doi:10.1109/TVCG.2011.220. 3

[KHGW07] Kuo B. Y.-L., Hentrich T., Good B. M., Wilkinson M. D.: Tag clouds for summarizing web search results. In *Proceedings of the 16th international conference on World Wide Web* (2007), WWW '07, pp. 1203–1204. doi:10.1145/1242572.1242766. 2

[KL07] Kaser O., Lemire D.: Tag-cloud drawing: Algorithms for cloud visualization. In *Workshop on Tagging and Metadata for Social Information Organization* (2007), WWW '07. 2

[KLKS10] Koh K., Lee B., Kim B., Seo J.: Maniwordle: Providing flexible control over wordle. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (2010), 1190 –1197. doi:10.1109/TVCG.2010.175. 2

[LRKC10] Lee B., Riche N. H., Karlson A. K., Carpendale S.: Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions Visualization and Computer Graphics 16*, 6 (2010), 1182 –1189. doi:10.1109/TVCG.2010.205. 2

[PNML08] Paulovich F. V., Nonato L. G., Minghim R., Levkowitz H.: Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics 14*, 3 (2008), 564–575. doi:10.1109/TVCG.2007.70443. 3, 8

[Sal91] Salton G.: Developments in automatic text retrieval. *Science 253* (1991), 974–980. doi:10.1126/science.253.5023.974. 3

[SB07] Stein B., Benteler F.: On the generalized box-drawing of trees - survey and new technology. In *International Conference on Knowledge Management* (2007), pp. 416–423. 2

[SKK00] Steinbach M., Karypis G., Kumar V.: A comparison of document clustering techniques. In *Workshop on Text Mining, ACM SIGKDD International Conference on Data Mining* (2000), pp. 109–110. 3

[SKK*08] Seifert C., Kump B., Kienreich W., Granitzer G., Granitzer M.: On the beauty and usability of tag clouds. In *12th International Conference Information Visualization* (2008), pp. 17–25. doi:10.1109/IV.2008.89. 2, 6, 7

[SOR*09] Strobelt H., Oelke D., Rohrdantz C., Stoffel A., Keim D. A., Deussen O.: Document cards: A top trumps visualization for documents. *IEEE Transactions Visualization and Computer Graphics 15*, 6 (2009), 1145–1152. doi:10.1109/TVCG.2009.139. 2

[SWL*10] Shi L., Wei F., Liu S., Tan L., Lian X., Zhou M. X.: Understanding text corpora with multiple facets. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2010), pp. 99–106. doi:10.1109/VAST.2010.5652931. 2

[VWF09] Viegas F. B., Wattenberg M., Feinberg J.: Participatory visualization with wordle. *IEEE Transactions Visualization and Computer Graphics 15*, 6 (2009), 1137 –1144. doi:10.1109/TVCG.2009.171. 2, 5

[WHS07] Wascher G., Haußner H., Schumann H.: An improved typology of cutting and packing problems. *European Journal of Operational Research 183*, 3 (2007), 1109–1130. doi:10.1016/j.ejor.2005.12.047. 4

[WPW*11] Wu Y., Provan T., Wei F., Liu S., Ma K.-L.: Semantic-preserving word clouds by seam carving. *Computer Graphics Forum 30*, 3 (2011), 741–750. doi:10.1111/j.1467-8659.2011.01923.x. 2