Story Gems: A Novel Approach for Semantic Aware Text Visualization

Naoya Oda *
Nihon University

Yosuke Onoue †
Nihon University

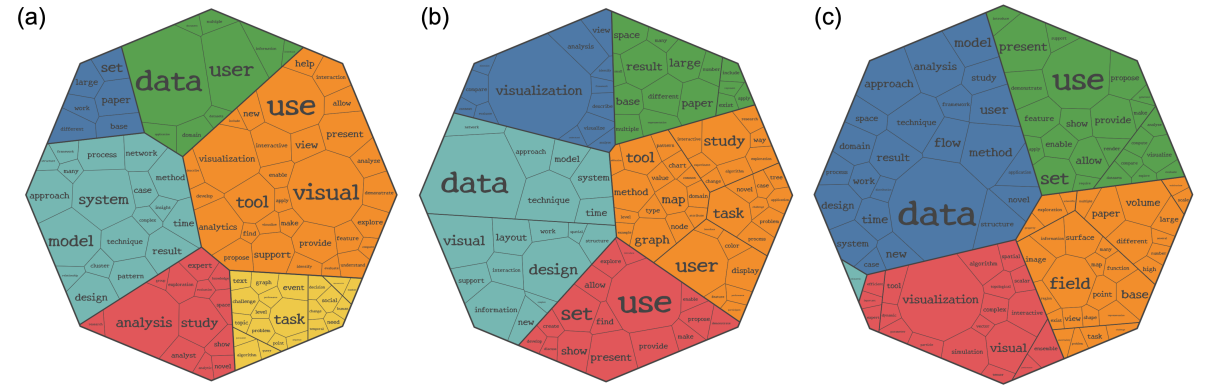

Figure 1: Visualization results of the proposed method Storgy Gem. (a), (b), and (c) are the visual summaries of article abstracts taken from VisPub dataset published in VAST, InfoVis, and SciVis conferences, respectively.

Abstract

Word cloud is a visualization method developed to help users quickly comprehend the overview and themes of an entire document. As a result of the initial word arrangement in current word clouds frequently being arbitrary, semantic links between words are not reflected. Several semantic-preserving word cloud techniques have been proposed; however, they have room for enhancement in spatial effectiveness. In this study, we suggest a revolutionary text visualization technique that preserves semantics. Story Jewel. Story Gem employs hierarchical clustering based on word network to reflect semantic relationships between words, and Voronoi treemap to place words with similar meanings close to each other within a specified region. From the application example to the VisPub dataset, it is verified that the presented method has the benefit that users can accurately recognize the visualization results due to the visibility of hierarchical data in Voronoi treemap.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Information visualization Human-centered computing—Visualization—Visualization techniques—Treemaps

1 Introduction

Word clouds are one type of text data visualization that are great for quickly summarizing and highlighting document content by grouping words from a text within a constrained space [5]. Most existing word cloud tools randomly place words, and not reflect semantic correlations between them in their position. To address these issuses, several semantic-preserving word cloud approaches have been proposed that reflect the relationship between words in the layout [3,6,7]. A popular method for creating semantically preserved word clouds places words based on word vectors generated through natural language processing.

There is a clear trade-off between classical word clouds and semantic-preserving word clouds. Although the arrangement of words in a traditional word cloud does not reflect their meaning, it offers good spatial efficiency and a lot of flexibility. Semantic-preserving word clouds tend to be sparse in word placement to reflect meaning in word placement. Furthermore, it is challenging to realize the function of laying words in an arbitrary shape, which tools such as Wordle have, in a semantic-preserving word cloud.

In this study, we offer a novel text visualization technique called Story Gem that manages to preserve semantics while placing words effectively. The proposed method arranges a hierarchical cluster structure of word networks created from input text within a particular region using Voronoi treemap [1]. Story Gem expresses the area of tiles in the Voronoi treemap in contrast to word clouds, which show the value of words in font size. This removes meaningless whitespace in semantic-preserving word clouds and improves spatial efficiency. Additionally, because Voronoi treemap visualize hierarchical data, the semantic proximity of words is maintained. We demonstrate the efficiency by applying the proposed method to the VisPub dataset.

2 Procedure of Story Gem

The procedure of Story Gem is mainly composed of the following three steps: (1) construction of a word network, (2) hierarchical clustering of the word network, and (3) visualization using a Voronoi treemap. Details for each step are provided below.

In the first step, Story Gem constructs a word network from input text by using several natural language processing methods At first, words are extracted by applying tokenization from the input text. Each word's relevance is calculated, and nonsensical stopwords and low-value words are eliminated.Next, we acquire the word vector and build a word network based on the vectors. In this paper, we use a pre-trained Word2Vec model word2vec-google-news-300 to acquire the word vectors. The frequency of terms in the input text divided by the vocabulary frequency of the model served as our measure of a word's value. The same effect as TF-IDF is obtained

*e-mail: oden6680@gmail.com
†e-mail: onoue.yousuke@nihon-u.ac.jp

by utilizing the word frequency of the model's vocabulary. Finally, the word network is constructed using k-nearest neighbor graph of the word vectors. In this study, we use cosine dissimilarity as a distance among words.

The second phase involves creating multi-level hierarchical clusters of words based on the word network's structure. We apply the Louvain method [2], which is known as a method of modularity-based community detection. Community identification is a node clustering technique applied to complicated network research. Modularity-based methods determine clusters so that fewer edges cross between clusters. In the Louvain method, the clustering result of one step is used as a coarsened network, and clustering is repeatedly conducted. Therefore, a hierarchical cluster structure of words is extracted from the input word network. After this phase, we have a tree with roots and leaves that represent words and their significance.

In the third step, place the words using the Voronoi treemap with the rooted tree constructed in the previous step as input. One type of treemap that displays hierarchical data with regions proportionate to numerical values is the Voronoi treemap. The Voronoi treemap uses Voronoi tessellation to insert tiles with numerical values in the designated area.. We use the significance of words for the numerical value expressed by the area of the tile. The font size of the words is determined so that it does not extend beyond the inscribed circle of the tile polygon. Presently, we use d3-voronoi-treemap (`https://github.com/Kcnarf/d3-voronoi-treemap`) as an implementation of Voronoi treemap. The uppermost cluster determines the hue of the tiles for the visual depiction. Nested clusters are denoted by the thickness of the lines separating the tiles.

3 Application Example

In this section, we present an application example of the Story Gem using VisPub dataset [4] to show the effectiveness. This dataset includes details on papers presented at IEEE VIS conferences between 1990 and 2021. We used the abstract texts of the articles published at each of the three conferences, VAST, InforVis, and SciVis. We empirically determined data processing parameters: $k = 18$ for k-nearest neighbors and extract top 100 words of significance value. Additionally, a regular octagon is selected as the outer region of the Voronoi treemap.

Figure reveals the visualization results of the Story Gem. We concentrate on the VAST results as an example. The result contains color coded 6 clusters: blue, green, red, orange, yellow, and light blue. In the orange region, we can observe that numerous words frequently used in visualization papers, such as "visualization" and "visual," are placed. In the green region, we can observe that several words commonly used to describe experiments, such as "data" and "user," are placed. In the light blue and red regions, words such as "system," "analysis," "model," and "study" are placed. Reading through these visualization results reveals that VAST is a journal about visualization with a focus on novel systems and applications for data analysis. This is demonstrated by the way that words are arranged, colors are assigned using clustering, and regions are divided using Voronoi treemaps. This is consistent with the features of VAST, and the visualization results derived from the abstracts of all papers submitted to VAST as input data were able to capture the outline of VAST, thus showing that the proposed method meets the expected effect to a certain degree. It is believed that Story Gem's exceptional spatial efficiency and proper semantic preservation made accurate text summary possible.

4 Conclusions

In this study, we proposed a novel semantic-preserving text visualization method Story Gem using hierarchical word clusters and Voronoi treemap. The presented method is composed of the following three steps: (1) creation of a word network appearing in input text, (2) using hierarchical clustering to showl semantically close words, and (3) applying Voronoi treemap to place semantically close words nearby. The visualization results shown in the application example generally agreed with the visualization results we had assumed in advance based on the data contents, and we can anticipate that this method will provide users good visualization results as a new semantic-preserving text visualization. The approach proposed in this study has the property that the visualization results vary greatly depending on the method of natural language processing applied to an input text. An interactive demo of the Story Gem is published in `https://storygem.vdslab.jp/`.

To improve visualization outcomes in the future, we are actively investigating the optimal natural language processing to apply to an input text. And the original Voronoi treemap enables tiles to be placed within arbitrary polygons, but presently tiles are placed only within convex polygons for implementation reasons. We plan to fix this by enhancing the implementation. Additionally, we intend to evaluate how valuable our approach is to users by comparing it to currently available word clouds and semantically preserved word clouds utilizing the visualization outcomes produced by our method.

References

[1] M. Balzer and O. Deussen. Voronoi treemaps. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 49–56. IEEE, 2005.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[3] M. A. Hearst, E. Pedersen, L. Patil, E. Lee, P. Laskowski, and S. Franconeri. An evaluation of semantically grouped word cloud designs. *IEEE transactions on visualization and computer graphics*, 26(9):2748–2761, 2019.

[4] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata. org: A metadata collection about ieee visualization (vis) publications. *IEEE transactions on visualization and computer graphics*, 23(9):2199–2206, 2016.

[5] F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *IEEE transactions on visualization and computer graphics*, 15(6):1137–1144, 2009.

[6] J. Wang, J. Zhao, S. Guo, C. North, and N. Ramakrishnan. Recloud: semantics-based word cloud visualization of user reviews. In *Graphics Interface 2014*, pp. 151–158. AK Peters/CRC Press, 2020.

[7] J. Xu, Y. Tao, and H. Lin. Semantic word cloud generation based on word embeddings. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 239–243. IEEE, 2016.