Visualisation of text corpora and search results

David Seywald
0931000
d.seywald@student.tugraz.at

706.720 Visual Analytics VU SS 2018
Graz University of Technology

May 2018

Table of Contents

Overview of Text Visualization Techniques

With the every growing amount of information contained in text document collections, it is imperative to allow the user to easily understand, search and explore the given information with the help of text visualization techniques.
In this Report, we want to give a short overview of the different text visualization techniques described in this book [1]. The authors group the various visualization techniques regarding their goal into five distinct categories:

- Visualizing document similarity
- Revealing content
- Visualizing sentiments and emotions of the text
- Exploring document corpus
- Analyzing various domain-specific rich-text corpus (Social media, news, emails etc.)

We will introduce and describe each category briefly and give an example of an interesting tool, which solves the specified task well.
An interactive list of over 400 different text visualization techniques, can be found here [2].
This Text Visualization Browser is one of the most complete collections for visualization techniques over the last decades.

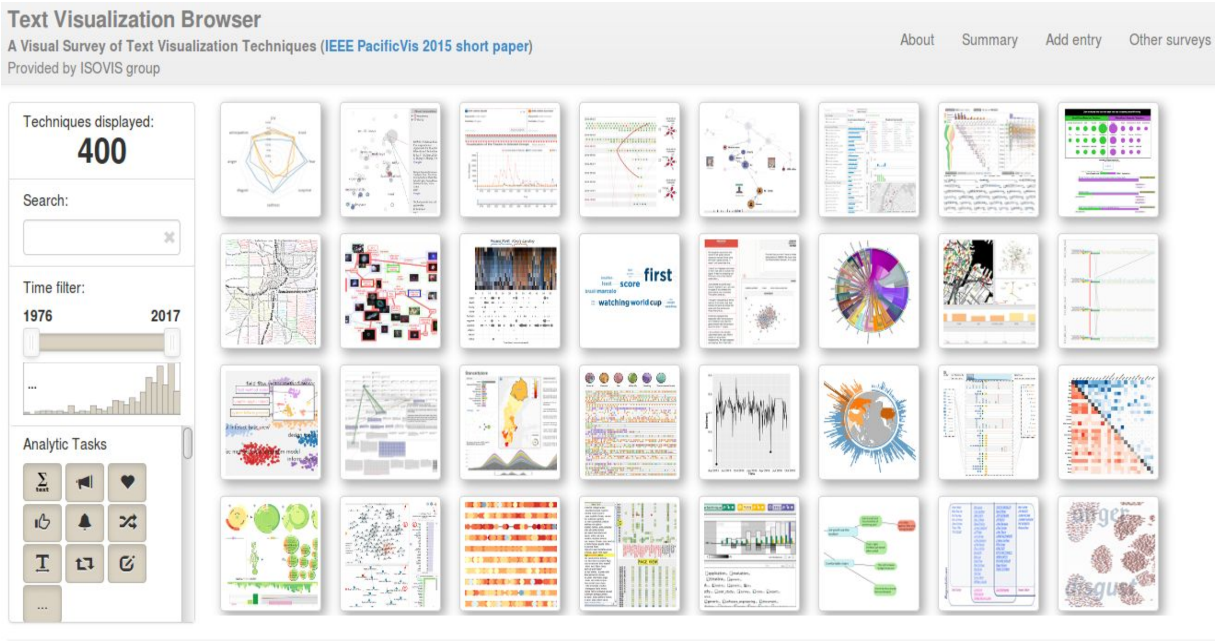

Figure: TextVis Browser, an interactive online browser of existing text visualization techniques

[1] "Overview of Text Visualization Techniques - Springer." https://www.springer.com/cda/content/document/cda_downloaddocument/9789462391857-c2.pdf?SGWID=0-0-45-1593357-p179855184. Aufgerufen am 17 Mai. 2018.

[2] "Text Visualization Browser." http://textvis.lnu.se/. Aufgerufen am 17 Mai. 2018.

Visualizing Document Similarity

One of the most common technique is to represent content similarities at the document level. Documents are usually represented as points in 2 or 3-dimensional space. The distance between the points describes the similarity of the corresponding documents. We can distinguish two different similarity techniques:

1. Projection-oriented
2. Semantic-oriented

Projection Oriented Techniques

In these techniques the content is split into a list of words, to create a n-dimensional feature vector, based on the most informative words. This is calculated by using "Term Frequency Inverse Document Frequency (TF-IDF)". Some common pre-processing tasks are also used to increase accuracy and performance, such as: removal of stopwords, stemming etc.
The resulting feature vector is then visualized by means of a dimensionality reduction algorithm. Common linear algorithms are "Principal Component Analysis (PCA)" or "Linear Discriminant Analysis (LDA)". Non-linear algorithms include "Multidimensional Scaling (MDS)" and "Locally Linear embedding (LLE)".

Semantic Oriented Techniques

These approaches concentrate on expressing similarities between documents based on their underlying topics. Common algorithms in this field include: "Probabilistic Latent Semantic Analysis (PLSA)", "Latent Dirichlet Allocation (LDA)", "Spherical Topic Model (SAM)", and "Non-Negative Matrix Factorization (NMF)".
Use Cases for this algorithms are mostly designed for analytical tasks and not for visualization, so their output may not be used directly. Still, compared to the projection oriented techniques, semantic based approaches can create more sensible results that user can easily understand.

Revealing content

The goal here is to visually represent the content of text documents. A single document can be summarized by two aspects: its content (words, sentences etc.) or its structure and metadata (average sentence length, word count, statistics etc.)
Interestings Projects for those tasks include the following:

- Wordle: A state-of-the-art Tag Cloud implementation.
- Document fingerprint: A heatmap where each cell represents a text block, with the feature value encoded as color (for example average sentence length).

Figure: Wordle visualization of a bag of words extracted from text data

Visualizing Sentiments and Emotions

Different techniques have been invented to visualize the change of sentiments over time. Some common examples include the sentiment analysis of twitter feeds regarding a specific topic over time. One basic approach is to simple draw a time-series diagram, where the changes in sentiments are shown on the y-axis. To address the shortcomings of this approach (why and how exactly did the sentiment changed), more sophisticated techniques are used. Interesting tools in this section are:

- EmotionWatch: Classify users’ emotional reactions to public events overtime in a radar diagram.
- Pulse: Visualize the topic/sentiment of large amounts of customer feedback, to examine opinions at a glance.

Document Exploration Techniques

The large part of our report focussed on document exploration. The goal here is to allow the user to navigate large document collections, discover patterns and gain useful information. Document exploration techniques can be split into four different categories.

Distortion Based Approaches

In this technique the focused part of the document is shown prominently in the view center, while the rest is distorted (blurred, smaller size, different shade). Early examples include “Document Lens” and “Data Mountain”.

Exploration Based on Document Similarity

Here full document collections are visualized in a multi coordinated view, based on their similarity. Users can navigate and interact with this collection by zooming and panning, resulting in different levels of details.Examples for such systems are “InfoSky” and “ForceSPIRE”

Hierarchical Document Exploration

Documents are clustered based on their similarity and hierarchically shown in a tree-like structure.

Search and Query Based Approaches

A classic approach, where the large amounts of documents are handled by reducing the displayed information via queries and filter. The result is ranked after various criterias and the users can navigate through different views to get exactly the information they need. Views are usually connected interactively, so changing on parameter also effects all the other views. This technique is called “Linking and Brushing”.

Conclusion

We briefly explained various text visualization techniques, based on their overall goal. Many different implementations already exists. For a comprehensive overview with over 400 examples, visit the Text Visualization Browser.

References

C. Nan and W. Cui, Introduction to Text Visualization, Atlantis Briefs in Artificial Intelligence 1, DOI 10.2991/978-94-6239-186-4_2

Viegas, F.B., Wattenberg, M., Feinberg, J.: Participatory visualization with wordle. IEEE Trans. Vis. Comput. Graph. 15(6), 1137–1144 (2009)

Kempter, R., Sintsova, V., Musat, C., Pu, P.: Emotionwatch: visualizing fine-grained emotions in event-related tweets. In: International AAAI Conference on Weblogs and Social Media (2014)

Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E.: Pulse: mining customer opinions from free text. In: Advances in Intelligent Data Analysis VI, pp. 121–132. Springer, Berlin (2005)

Andrews, K., Kienreich, W., Sabol, V., Becker, J., Droschl, G., Kappe, F., Granitzer, M., Auer, P., Tochtermann, K.: The infosky visual explorer: exploiting hierarchical structure and document similarities. Inf. Vis. 1(3–4), 166–181 (2002)

Endert, A., Fiaux, P., North, C.: Semantic interaction for visual text analytics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 473–482. ACM (2012)

Keim, D., et al.: Information visualization and visual data mining. IEEE Trans. Vis. Comput. Graph. 8(1), 1–8 (2002)

Keim, D., Oelke, D., et al.: Literature fingerprinting: a new method for visual literary analysis. In: IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007, pp. 115–122. IEEE (2007)