Advanced Text Visualizations

Hendrik Strobelt (hendrik.strobelt@uni-konstanz.de)
Advisor: Prof. Dr. Oliver Deussen

Nowadays, large document collections, such as research paper corpora and news feeds, grow at high rate. Many of these documents contain text and images for describing facts, methods, or telling stories. It is an exhaustive task for a user to get an overview of a single document and especially of a larger collection. To overcome this problem, it is common to represent a document in a different way. For instance, search engines usually show the title of a document together with a small context of the query terms. With this representation a user has to read only a portion of the text and is focused on the relevant parts of the documents, which efficiently allows him or her to differentiate between relevant and non-relevant documents. While this representation is efficient to browse through search results, it is not capable to give a quick overview of a whole document collection.

The work is concentrated on finding such a representation that aims at the task of recognition and provide an information rich compact view on a document for the task of exploration. The challenges are on one hand side the extraction of data from various sources like PDF files and webpages. On the other hand, for reducing the dataset to a visualizable subset it is needed to extract from these fulltext sources and image sources these items that are good representatives. These two challenges are getting solved in a cooperation with text mining specialists of the Graduiertenkolleg. After acquiring and filtering the text and image data, the key aspect is to find a suitable visualization.

The visualization should show the documents content in an one-view representation. This allows the use of this view as visual nodes in networks, or as node for visual clustering, etc. Supporting this task, the views should be discriminative enough to allow fast differentiation even at small scale. They should also offer interaction possibilities for deeper views inside the documents or represent additional features. In a first approach the Document Cards (DC) were invented to create such a view on documents. Each DC shows some well chosen and weighted images, important key terms, and additional static text features (like title and author names).

For a future work the visual aspect of semantical zooming these document visualizations will be on the focus of the research. The idea of simple context switching on different level of details will be replaced by a smooth transformation that allows a trackable view on every zoom level and therefore takes respect of the limited tracking capability of humans recognition. For that goal, good layout algorithms are needed, that provide a compact rectangle image packing that remains stable during insertion and deleting rectangles. Furthermore the text layout will be in special focus since it underlies stricter constraints while zooming, like readability.