Efficient Visualization of Large Text Corpora

Marko Grobelnik, Dunja Mladenić
marko.grobelnik@ijs.si, dunja.mladenic@ijs.si
Jozef Stefan Institute, Jamova 39 Ljubljana, Slovenia

Abstract

Visualization is one of the important ways on how to deal with large amounts of textual data. Most frequent application of text visualization techniques is particular in cases when one needs to understand or to explain the structure and nature of large quantity of typically unlabeled and poorly structured textual data in the form of documents.

The usual approach when dealing with text for visualization is first to transform the text data into some form of high dimensional data and in the second step to carry out some kind of dimensionality reduction down to two or three dimensions that allows to graphically visualize the data. There are several (but not too many) approaches and techniques offering different insights into the text data like: showing similarity structure of documents in the corpora (e.g. WebSOM, ThemeScape), showing time line or topic development through time in the corpora (e.g. ThemeRiver), showing frequent words and phrases relationships between them (Pajek), etc.

One of the most important issues when dealing with visualization techniques is scalability of the approach to enable processing of very large amounts of the data. In this paper, our contributions are two procedures for text visualization working in linear time and space complexity.

The first procedure is a combination of the K-Means clustering procedure and a technique for nice graph drawing. The idea is first to build certain number of document clusters (with K-Means procedure), which are in the second step transformed into the graph structure where more similar clusters are connected and bound more tightly. The third step performs one sort of multidimensional scaling procedure by aesthetically drawing of the graph. Each node in the graph represents the set of similar documents represented by the most relevant and distinguishing keywords denoting the topic of the documents.

The second procedure performs hierarchical K-Means clustering procedure producing a hierarchy of document clusters. In the next step the hierarchy is drawn into the two-dimensional area split accordingly to the hierarchy splits. Like in the first approach, each cluster (group of documents) in the hierarchy is represented by the set of the most relevant keywords.

Both approaches will be demonstrated on the number of examples visualizing e.g. Reuters text corpora (over 800k documents) and various web-sites.