The Infomat Interface - 
A Document Clustering Exploration Interface

Magnus Rosell

There is a lot of valuable information hidden in unstructured text. Think for instance of the free text answers from medical questionnaires, which could contain reasons and/or possible cures for diseases. Most of such answers are never thoroughly investigated since it is too time consuming to do manually. Automatic tools for this could be very valuable.

Document clustering groups similar documents together. The result is often presented in text: some important words and documents are given for each cluster. This is often good to understand what a cluster is about, but it may be difficult to get an overview of the whole result.

The task is to build a graphical user interface and visualization tool that help the user (for example a researcher at Karolinska Institutet) in investigating document clustering results. The results of a document clustering system will be the input.

In Information Retrieval (the field of search engine and the like) the relations between documents and words are sometimes modelled by the document-by-term matrix\(^1\). Each element in this matrix is a weight that indicate how important a particular word is to describe the content of a particular document.

Figure 1 gives a rough picture of how the interface will look like. The intensity of each pixel is proportional to the weight of the matrix elements it represents. It should be possible to resort the documents and words in different ways.

To highlight different aspects the interface should give the possibility to color different parts of the matrix. For instance, if one has a manual categorization of the same set of texts, the documents belonging to the same category could be given a certain color.

Further, a “lens” should be implemented. The words and documents it is placed over should be magnified so that the words and document names appear at the edges of the matrix. This gives the user the possibility to zoom in on interesting looking parts of the matrix.

Care must be taken to make the program fast. Recalculations and redrawing must be performed fast enough for the tool to be used interactively. The best implementation will be used within the research project Infomat, see:

http://www.nada.kth.se/theory/projects/infomat/

\(^1\)Note that there is no consideration taken to order between words and/or documents here.
Figure 1: A simple view of the document-by-term matrix. A set of 2500 documents (rows) with 7096 words (columns) has been clustered to five clusters (separated by black horizontal lines) with the K-Means algorithm. The matrix has been rounded to a 300 by 300 pixels picture. The documents are sorted within each cluster according to similarity to the cluster, the most similar at the top of the cluster. The words are also sorted. Words with the highest average weightning in the top cluster are displayed as the leftmost columns, and so on. The words are also sorted within these groups; higher weighting further to the left.