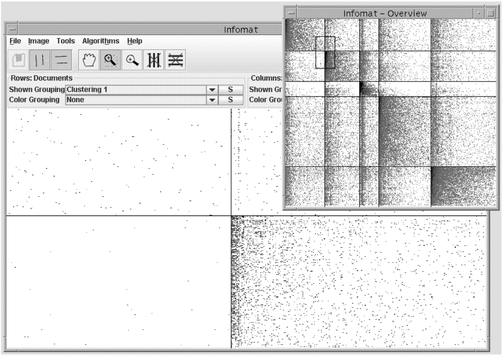
Infomat
A Vector Space Visualization Tool

Magnus Rosell
KTH CSC
100 44 Stockholm
Sweden
rosell@csc.kth.se

Infomat is a vector space visualization tool aimed at Information Retrieval (IR) and text clustering in particular. However, it could be used in many areas of language technology, as well as in other fields when information can be stored in a matrix (Infomat – information matrix).

As an example we give an IR matrix where rows represent texts and columns words, see Figure 1. In each matrix element the tf*idf weight for the word in the text is stored. Similarity between texts is calculated with the cosine measure.



Shown Grouping Clustering 1
Color Grouping None

Figure 1: Infomat. The upper right window is the overview window, displaying the whole matrix. The small rectangle in it indicates the part that is displayed in the main window. Using K-Means 2500 Swedish newspaper articles have been clustered to five clusters along the rows. The columns represent 5663 words, clustered to five clusters relative to the row (article) clusters. Hence the diagonal pattern in the overview.

In Infomat the matrix is presented as a scatter plot, where the opacity of each pixel is proportional to the weight of the corresponding matrix element(s). When the mouse pointer is placed over a pixel textual information about its content is presented.

When the matrix is larger than the picture it is compressed and each pixel presents the average value of the corresponding matrix elements. This can be justified when the objects of adjacent rows and columns are related (have high similarity). One way of obtaining such relatedness is through clustering. Infomat provides basic clustering functionality.

Infomat has many functions. Among other things it is possible to zoom in and out of the matrix. To visualize several groupings (clusterings, categorizations, etc.) at the same time it is possible to color objects belonging to different groups in different colors, both in rows and columns.

Many existing IR visualization methods calculate the similarity between all objects and project this relationship down to two or three dimensions (Baeza-Yates and Ribeiro-Neto, 1999). Such methods do not usually give much information as to why objects are deemed similar. In Infomat similarity between adjacent rows and columns appear as patterns, reflecting the distributional definition of similarity.

Infomat is developed in Java and uses an xml-format for reading and writing matrixes. It is freely available[1] together with more information.

References

R. Baeza-Yates and B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley.

[1] http://www.csc.kth.se/tcs/humanlang/