

# Approaches to Document Classification and Visualisation

Johan Hagman, Ralf Steinberger, Domenico Perrotta, Aristide Varfis

European Commission - Joint Research Centre (JRC)

Institute for Systems, Informatics and Safety (ISIS)

System Analysis and Information Assessment Unit (SAIA)

T.P. 361, 21020 Ispra (VA), Italy

{johan.hagman/ralf.steinberger/domenico.perrotta/aristide.varfis}@jrc.it

## Abstract

In this short paper we present two clustering and visualisation techniques for document collections which have been developed at the *Joint Research Centre* to support specific users within the *European Commission*. The visualisation tools will be part of a complex document retrieval, information extraction and visualisation system.

## 1 Introduction

Like any other organisation, the European Commission (EC) has a need to monitor activities and events in its fields of interest. To this end, the *Joint Research Centre* (JRC) is building a document retrieval, information extraction and visualisation system which has the goal of making large amounts of relevant information from the internet and the EC's intranet accessible in an efficient and easily intelligible manner.

Both visualisation methods described in the following sections organise large collections of documents (or other items) in a two-dimensional space (a 'map') where similar documents are placed close to each other. The purpose of this is to give users a quick overview of the collection and to retrieve documents which are similar to a chosen one. We explain both approaches on the basis of a small sample of 260 English documents. As with all statistical methods, a larger sample will improve the accuracy of the system.

In addition to these visualisation techniques, relevant components of this system comprise a tool to retrieve documents safely from the internet, a text pre-processing tool which also extracts some basic entities such as dates, countries and legal references, a bigram-based language identifier, a lemmatiser, a keyword identification tool, a subject domain identifier and a word/document clustering tool. The document visualisation techniques are applied at the end of a series of retrieval, normalisation and information extraction steps. Normalisation also includes the translation of non-French texts into French, using the EC's machine translation system *Systran*, as well as the

recognition of multi-word terms using the EC's *Text Analysis Tool*, which is part of the *EURAMIS* workbench [T & T, 1998].

## 2 WEBSOM

The first visualisation technique is a customised version of a neural network approach called WEBSOM. WEBSOM is a method which has been developed by the Neural Network Research Centre of Helsinki University of Technology ([Kohonen et al., 1996]) and which organises document collections and presents them in two-dimensional space. The method owes its name to its original use, which was a successful large-scale implementation of the system using internet (web) documents, on the one hand, and, on the other hand, to the fact that it is based on two successive runs of the connectionist Self Organising Map (SOM) model. The SOM algorithm is considered one of the most flexible algorithms amongst

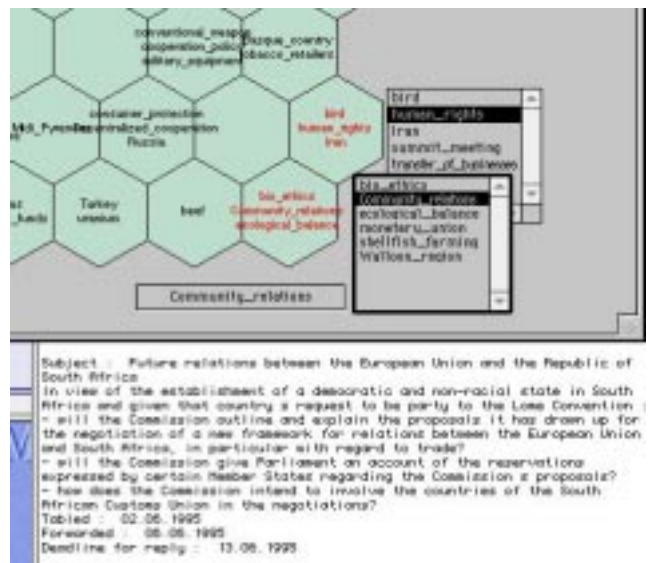


Figure 1 Snapshot of the lower right part of a document map, plus the full text of the 'Community relations' document.

many that map high-dimensional data such as document vectors onto a much more tractable two-dimensional representation.

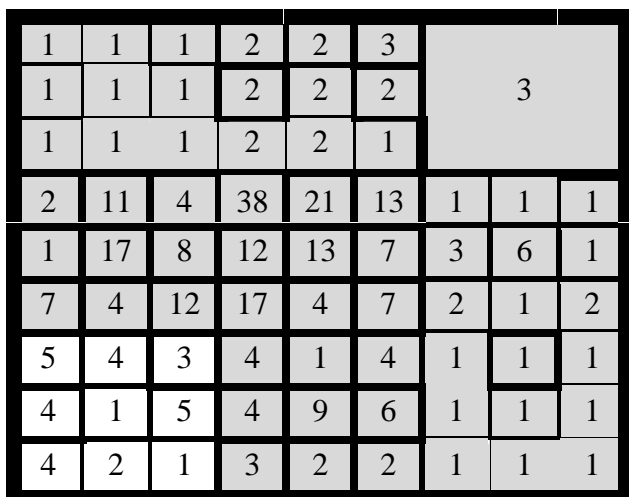
Loosely speaking, due to the SOM organisation properties, documents which will be attached to the same or to neighbouring cells in the 'document map', will be expected to be 'similar'. Figure 1 displays the bottom-right region of a document map with 18x18 cells, which was trained on the small set of 260 documents. Each cell shows a maximum of three document names. Clicking on one cell opens a scrolling list window, and selecting a document name in the list lets its full text be available.

### 3 Cluster analysis yielding doc- or word-maps

The alternative technique we offer users to get an overview over collections of documents is based on cluster analysis. Besides documents, items such as indexing words can be visualised. Here we choose to illustrate the procedure with the indexing words because their organisation into clusters is easier to understand using one's common thesaural knowledge.

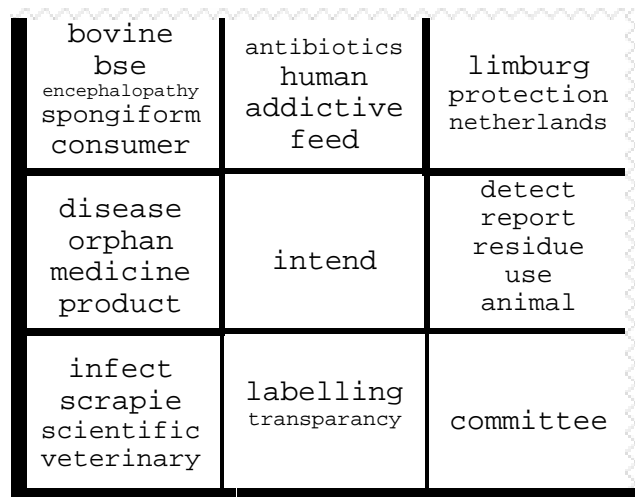
Once the textual data has passed through the modules of pre-processing, language recognition, translation, lemmatisation, and keyword identification, we have a table where each document is represented by a set of (possibly weighted) indexing terms. On the basis of these tables we calculate two similarity matrices: the wordXword matrix and the docXdoc matrix (see, e.g. [Oakes, 1998: 110-120]). From the similarity matrices we create the corresponding dendrograms. The algorithm is binary, hierarchical, agglomerative, and uses the average linkage between the documents/words. Except for the similarity calculus, the procedure is virtually the same for both dendrograms.

In order to visualise how the indexing words or the documents relate to each other in a more compact way



**Figure 2** Mapping of a dendrogram onto a 9x9 grid where the cell walls indicate the inter-subtree similarities. Numbers indicate subtree sizes. Non-shaded part is zoomed into by Figure 3.

than the large "1½-dimensional" dendrograms permit, we map them onto a two-dimensional grid. The procedure is to first cut up the trees into nine subtrees (by starting from the root and dissolving the weakest nodes until there are 9 subtrees) and then to distribute these within a 3x3 grid in such a way that more related subtrees lie closer to each other. Then we continue this division and divide those of the nine subtrees which contain at least nine items into nine further subtrees (subsubtrees). When calculating the best cell for each subsubtree in order to have similar cells close to each other, we consider all eight subtree-internal and the closest neighbouring subtree-external cells/subsubtrees. The result of this is shown in Figure 2 where we also visualise the similarity between each neighbouring subsubtree pair by the thickness of the cell walls: the thinner the wall, the more similar the items. The numbers indicate the number of documents associated to each cell. Figure 3 zooms in on the lower left corner of Figure 2 and shows the keywords contained in its cells.



**Figure 3** Detail of Fig. 2 showing how 29 of the 321 indexing terms used in a small test run are optimally mapped onto two dimensions.

For larger document collections, this 9x9 grid must be collapsed further and the cells should be made 'clickable' to make their underlying subtree pop up in an interactive session of data exploration or information retrieval.

### References

- [Kohonen et al., 1996]. T. Kohonen et al. Very large two-level SOM for the browsing of newsgroups. In Proceedings of ICANN'98, 1996. Springer, Berlin.
- [Oakes, 1998]. Michael P. Oakes, *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press, Edinburgh, 1998.
- [T & T, 1998]. *Terminology and Translation 1.1998 – a Journal of the Language Services of the European Institutions*, Office des publications officielles des Communautés européennes, Luxembourg, 1998.