

Supporting Information Management in Digital Libraries with Map-based Interfaces

Rudolf Mayer, Angela Roiger and Andreas Rauber

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Vienna, Austria
mayer@ifs.tuwien.ac.at, angela@roiger.at, rauber@ifs.tuwien.ac.at

Abstract. The *Self-Organising Map* (SOM) has been proposed as an interface for exploring Digital Libraries, in addition to conventional search and browsing. With advanced visualisations uncovering the contents and its structure, and advanced interaction modes as zooming, panning and area selection, the SOM becomes a feasible alternative to classical interfaces. However, there are still shortcomings in helping the user to understand the map – there are insufficient methods developed for describing the map to support the user in the analysis of the map contents. In this paper, we present recent work in assisting the user in exploring the map by automatically describing maps using advanced labelling and summarisation of map regions.

Key words: Self-Organising Map, Interface, Summarisation, Clustering

1 Introduction

The Self-Organising Map (SOM) [1] is an unsupervised neural network model that provides a topology preserving mapping from a high-dimensional input space to a low, often two-dimensional, output space. The output space consists of a grid of units, each associated with a weight vector of the same dimensionality as the input space. During the training process, the weight vectors are adapted to describe the input space as close as possible, arranging topically related inputs close to each other.

The SOM has been used in several applications to automatically organise documents in a Digital Library by their content. Examples are text, as in the WEBSOM project [2], the SOMLib Digital Library system [3], or in a map of news [4], music in the SOMeJB system [5], or pictures in the PicSOM project [6]. Advanced visualisations and interaction possibilities allow the user to fully exploit the potential of the SOM. An extension of the SOMLib system e.g. realises the symbiosis of traditional information retrieval using a search or list browsing interface with the explorative approach of the Self-Organising Map by providing a plug-in to the popular open-source Digital Library System *Greenstone 3* [7]. The user can exploit all the functionalities provided by Greenstone, and can benefit from the wealth of additional information the SOM mapping provides

about the cluster structure of the documents matching the query results, and the whole collection itself.

However, we still lack techniques to adequately help the user analysing the contents of the map. For large maps, containing several tens of thousands of documents on various topics, it becomes increasingly difficult to analyse the map. In this paper, we give an overview of existing uses of the Self-Organising Map in Digital Libraries and techniques to explore and interact with the map. Furthermore, we present recent work in automatically describing regions in the map, using clustering methods to identify topical areas and selecting representative labels and summarising the content for those regions.

2 Describing the Self-Organising Map Regions

In this section we present our work on identifying and describing regions in the SOM. As the SOM does not generate a partitioning of the map into separate clusters, we apply a clustering algorithm on the weight vectors of the units to identify the regions (Section 2.1). We then extract semantic labels for those regions (Section 2.2), that assist the user in getting a first glance overview of the contents of the map. To further support the analysis, we provide summarisation of documents using Automatic Text Summarisation methods (Section 2.3).

2.1 Clustering

Clustering is an unsupervised process of finding natural groupings amongst unlabelled objects. We cluster the units of a SOM using an agglomerative, hierarchical clustering algorithm on the weight vectors. In the beginning of this algorithm, every unit lies in its own cluster. In each subsequent step, the two nearest clusters are merged, until finally only one cluster remains. Specifically, we use *Ward's linkage* as one of the most performant within the linkage clustering families, where the distance of each pair of clusters is defined by the increase in the 'error sum of squares' if the two clusters are to be combined. The result of the algorithm is a hierarchy of clusters which the user can browse through. Increasing the number of displayed clusters means splitting existing clusters into two new ones, while reducing the number of clusters is achieved by merging two clusters into one. This is advantageous over a non-hierarchical clustering algorithm, where changing the number of clusters might completely change the layout of clusters. Moreover, hierarchical clustering allows us to display multiple layers with a different number of clusters at the same time.

2.2 Labelling Regions

To assist the user in interpreting the regions of the SOM, we automatically generate labels for the clusters we identified previously. The cluster labels are based on the unit labels generated by the LabelSOM method [8], which assigns labels to the units of the SOM describing the features of the data points mapped

onto the respective unit. This is done by utilising the *quantisation error* of the vector elements, i.e. the sum of the distances for a feature between the unit's weight vector and all the input vectors mapped onto this unit. A low quantisation error characterises a feature that is similar in all input vectors to the weight vector. Thus the assumption is made that this feature describes the unit well. If the input vector contains a lot of attributes which are non-existent and therefore have the value of 0, those attributes often also have a quantisation error of almost 0 for a unit. However, such features are in most cases not appropriate for labelling the unit, since they would describe what the unit does not contain. Therefore, we require vector elements to additionally have a mean value above a defined threshold. To choose a label for a region, we consider all the unit labels present in that cluster. We chose to determine the cluster labels based upon the unit labels as they are already a selection of features describing the contents of each unit. This method is faster in computation than checking all possible features. The user can specify whether he prefers labels based upon a low average quantisation error, a high mean value, or a combination of both. Utilising the properties of the hierarchical clustering we can also display two or more different levels of labels, some being more global, some being more local.

In the visualisation of the SOM, the labels are placed in the centroid of the cluster, which may result in some overlapping labels. To achieve a clear arrangement, labels can be manually moved on the map, or adjusted in their size and rotation. For some labels, it might be useful to edit their text, for example if the label text is only a word stem as in the experiment described below.

2.3 Region Summarisation

Even though labelling the map regions assists the user in quickly getting a coarse overview of the topics, labels can still be ambiguous or not conveying enough information. Therefore, we also employ Automatic Text Summarisation [9] methods. We provide the user with summaries of single documents (based on extraction methods), however, the main focus is to assist the user in quickly analysing the contents of the Digital Library by providing summaries of the previously identified regions using multi-document summarisation. The application allows the user to select whole regions, or manually any other rectangular shape or units along a path. From the documents on those units, the user can choose from several different summarisation algorithms using different weighting schemes to determine the importance of sentences for the summaries. Further, the user can specify the desired length of the summary, measured in percent of the original sentences. Thus, a somewhat more concise description of the topical areas is provided to the user.

3 Experiments

The following experiments were performed using the 20 newsgroups data set (<http://people.csail.mit.edu/jrennie/20Newsgroups>). It consists of 1000

newsgroup postings for each of its 20 different newsgroups, such as *alt.atheism* and *comp.sys.mac.hardware*. We used a bag-of-words indexing and Porter’s stemming. Features for the input vectors were selected according to their document frequency, and the weights are computed using a standard $tf \times idf$ weighting scheme. This resulted in a 3151 dimensional feature vector for each document, from which the maps were trained. The specific map we will use in the remainder of this paper to illustrate our results is 75x55 units in size.

3.1 Labelling Regions

In our application it is possible to explore the clustered SOM interactively: to view the different levels of clustering and to zoom into the map to view the single postings. The user can browse the clustering levels either viewing only the cluster borders, or highlighting each cluster in a different colour. The former is shown in Figure 1, illustrating the steps from one to eight clusters. There is a special cluster in the lower right-hand corner with the a label also used on other clusters – ‘god’ in the fourth step, and ‘gun’ in the steps five to seven. This cluster is, however, not a separate one - when viewing the clustering with colours, it becomes apparent that this area is part of the disjoint clusters ‘god’ and ‘gun’, respectively, in the upper part of the map.

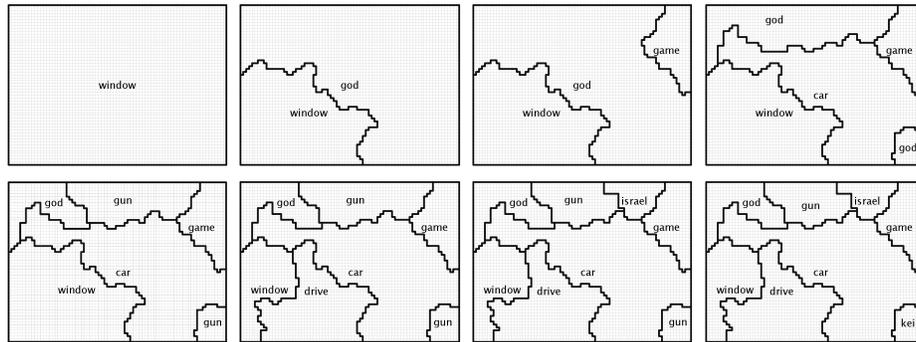
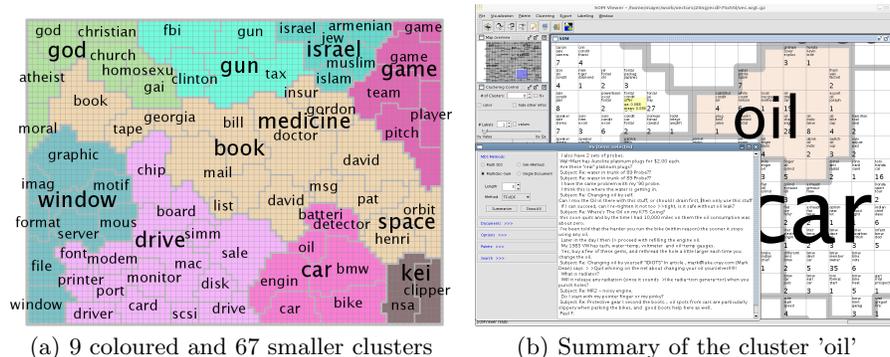


Fig. 1. 1 – 8 Clusters With Labels

Figure 2(a) shows an already labelled map, holding nine clusters with larger labels, and in addition 67 clusters with smaller labels. Again, there is one disjoint cluster, labelled ‘david’. As a result of the stemming, the labels are sometimes not complete words that could be expanded to their original form.

In the top right-hand corner is a large cluster labelled ‘game’ containing most postings from the two sports related newsgroups. The large cluster next to it labelled ‘israel’ contains mainly postings from *talk.politics.mideast*. In the upper left-hand corner there is a cluster labelled ‘god’ containing all the newsgroups dealing with religion, i.e. *alt.atheism*, *soc.religion.christian* and *talk.religion.misc*.



It is interesting to note, that, in contrast to the newsgroup hierarchy where these groups lie in three different top level hierarchies, they are combined into one cluster here.

The large cluster in the middle labelled 'book' contains many of the smaller clusters of which only a few have meaningful labels. The clusters labelled 'insur' and 'doctor' suggest that they contain postings from the sci.med newsgroup and the cluster label 'orbit' relates to the newsgroup sci.space. The labels containing names such as 'gordon', 'david' or 'bill' do not help in identifying the underlying topics in those areas of the map. However, names cannot be easily automatically removed, as some common names as Mark or Bill are also a verb or noun respectively. Furthermore names can sometimes be useful labels, for example if they refer to a famous person. The small cluster labelled 'drive' lies in the cluster with the hardware topics but also directly next to the cluster labelled 'car'. It implies that in this area lies an transition of the word *drive* being used in the meaning of *hard disk drive* or in the meaning of *to drive a car*.

To enhance the comprehensibility of the map, the user can manually edit labels, e.g. add word endings, or improve the labels of areas that have diverse topics. For example the cluster automatically labelled 'car' could be extended to 'car & bike' to point out both newsgroups contained in this cluster. The cluster 'gun' containing the newsgroups talk.politics.guns and parts of talk.politics.misc and talk.politics.mideast could be adapted to 'politics'.

3.2 Region Summarisation

Figure 2(b) shows the summarisation of one of the regions in map, namely the cluster labelled 'oil'. The lower-left part of the interface shows the summarisation module, which allows the user to select a summarisation method, and the desired length of the summary. In our example, we use a multi-document summarisation extracting sentences considering their importance for the whole

collection of documents selected, and chose 3% of the selected documents as desired summarisation length.

4 Conclusion

In this paper we presented the usage of the SOM as an interface to Digital Libraries. On top of this well-known approach, we presented methods to assist the user in interacting with the map. We employ clustering of the SOM to reveal hierarchical structures to provide a rough overview of the structure of the data. The clustering identifies regions, which we describe on the one hand by single descriptive words extracted from the document contents, and secondly by applying automatic text summarisation techniques to generate extracts of the contents. All methods are integrated into a single application that provides additional features such as visualisations and advanced interaction via zooming and panning, and selection of arbitrary regions of the map. With these tools available, the user can be greatly assisted in analysing the map and getting a quick overview of the contents of the Digital Library itself.

References

1. Kohonen, T.: Self-Organizing Maps. Volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg (1995)
2. Kohonen, T., Kaski, S., Lagus, K., Salojrvi, J., Paatero, V., Saarela, A.: Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery* **11** (2000) 574–585
3. Rauber, A., Merkl, D.: The SOMLib digital library system. In: *European Conference on Digital Libraries (ECDL 1999)*, Paris, France, Springer (1999) 323–342
4. Ong, T.H., Chen, H., Sung, W., Zhu, B.: Newsmap: a knowledge map for online news. *Decision Support Systems* **39** (2005) 583–597
5. Rauber, A., Frühwirth, M.: Automatically analyzing and organizing music archives. In: *Proceedings of the 5. European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, Darmstadt, Germany, Springer (2001)
6. Laaksonen, J., Koskela, M., Laakso, S., Oja, E.: PicSOM-content-based image retrieval with self-organizing maps. *Pattern Recogn. Lett.* **21** (2000) 1199–1207
7. Mayer, R., Rauber, A.: Adding SOMLib capabilities to the Greenstone Digital LibrarySystem. In: *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL)*, Springer (2006) 486–489
8. Rauber, A., Merkl, D.: Automatic labeling of Self-Organizing Maps for Information Retrieval. *Journal of Systems Research and Inf. Systems (JSRIS)* **10** (2001) 23–45
9. Mani, I.: *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA (1999)