Representation of Document Archives for Interactive Exploration

Dieter Merkl, Andreas Rauber
Institut für Softwaretechnik, Technische Universität Wien
Favoritenstraße 9–11/188, A–1040 Wien, Austria
www.ifs.tuwien.ac.at/∼dieter www.ifs.tuwien.ac.at/∼andi

Abstract

Today's information age may be characterized by constant massive production and dissemination of written information. More powerful tools for exploring, searching, and organizing the available mass of information are needed to cope with this situation. In this context the map metaphor for displaying the contents of a document archive in a two-dimensional display has gained increased interest. In particular, we rely on self-organizing maps, which produce a map of the document space after their training process.

From geography, however, it is known that maps are not always the best way to represent information spaces. For most applications it is better to provide a hierarchical view of the underlying data collection in form of an atlas, where, starting from a map representing the complete data collection, different regions are shown at finer levels of granularity. Using an atlas, the user can easily "zoom" into regions of particular interest while still having general maps for overall orientation.

We show that a similar display can be obtained by using the Growing Hierarchical Self-Organizing Map (GHSOM) to represent the contents of a document archive. This neural network model has an adaptive layered architecture where each layer consists of a number of individual self-organizing maps. By this, the contents of the text archive may be represented at arbitrary detail while still having the general maps available for global orientation.

1 Introduction

During the last years we have witnessed an uninterrupted rise of the amount of information available in electronic form. While the size and availability of electronic information has changed a lot, ways for representing and interacting with those collections could not keep pace. Searching requires users to define their queries in expressions based on boolean logic, specifying large numbers of keywords, synonyms and antinyms, requiring both knowledge of the problem domain as well as basic query formulation experience. Results of queries are usually presented as long lists of retrieved documents sorted following some ranking criteria, with the large overall number of documents retrieved usually inhibiting efficient search. Information on the documents retrieved from a collection is at the most presented as a rather long textual description of the available metadata.

While these systems allow the location of documents based on well-defined queries, the aspect of exploration of document archives has found only limited attention. The users should benefit particularly from clustering techniques that uncover similar documents and bring these similarities to the user's attention. One step in this direction is the self-organizing map (SOM), a popular unsupervised neural network, providing a map-based representation of a document archive. In such a representation, documents on similar topics are located next to each other. The obvious benefit for the user is that navigation in the document archive is similar to the well-known task of navigating in a geographical map. This allows the user to obtain an overview of the topics covered in a collection and their importance with respect to the amount of information present in each topical section.

Yet, most of the research work aims at providing one single map representation for the complete document archive. As a consequence, hierarchical relations between documents are hidden in the display. Moreover, it is only natural that with increasing size of the document archive the maps for representing the archive grow larger, thus leading to problems for the user in finding proper orientation within the map.

We believe that the representation of hierarchical document relations is vital for the usefulness of map-based document archive visualization approaches. In much the same way as we are showing the world on different pages in an atlas where each page contains a map showing some portion of the world at some specific resolution, we suggest to use a kind of atlas for document space representation. A page of this atlas shows a portion of the library at some resolution while omitting other parts of the library. As long as general maps that provide an overview of the whole library are available, the user can find her way along the library chosing maps that provide the most detailed view of the area of particular interest.

In this paper we argue in favor of establishing a hierarchical organization of the document space based on an unsupervised neural network. We are currently evaluating strategies to address these issues within the framework of our SOMLib project [14]. The SOMLib project is based on unsupervised neural network technology for the task of document archive organization. In particular, we rely on self-organizing maps [5] and variants thereof for document clustering. In order to detect hierarchical document relations, we proposed a novel neural network architecture, the *growing hierarchical self-organizing map* [2]. The distinctive feature of this model is its problem dependent architecture which develops during the unsupervised learning process. Starting from a rather small high-level SOM, which provides a coarse overview of the various topics present in a document collection, subsequent layers are added where necessary to display a finer subdivision of topics. Each map in turn grows in size until it represents its topic to a sufficient degree of granularity. Since usually not all topics are present equally strong in a collection, this leads to an unbalanced hierarchy, assigning more "map-space" to topics that are more prominent in a given collection. This allows the user to approach and intuitively browse a document collection in a way similar to conventional libraries.

The training process results in a hierarchical arrangement of the document collection where self-organizing maps from higher layers of the hierarchy are used to represent the overall organizational principles of the document archive while maps from lower layers of the hierarchy are used to provide fine-grained distinction between individual documents. Such an organization thus comes close to what we would usually expect from conventional libraries. As an important benefit from the unsupervised training process we have to note that the library organization is derived solely from the document representation. No semantic labeling such as labels of subject matters and the like is necessary.

First experiments with this model on a large document archive of newspaper articles demonstrated its general feasibility in uncovering hierarchical document relations [12]. For the user this approach has the benefit of enabling an explorative access to large document archives where zooming into topics of interest is realized in an easy and intuitive fashion. The intuitive access to document archives is further provided by making use of an interface incorporating metaphor graphics for meta data associated with the documents, e.g. size of the document, time of last access, frequency of access [11].

The remainder of this paper is structured as follows: Section 2 provides a brief introduction to the architecture and training process of the self-organizing map. Section 3 introduces the Growing Hierarchical Self-Organizing Map. The data set used for our experiments is presented in Section 4, with experimental results being discussed in detail in Section 5. Section 6 then provides an overview of the other modules of the SOMLib digital library system which allow the integration of distributed library maps, automatic labeling of the topical clusters and a metaphor-graphical representation of the documents, followed by some conclusions in Section 7.

2 Self-organizing maps

The self-organizing map [4, 5] is one of the most prominent artificial neural network models adhering to the unsupervised learning paradigm. It provides a mapping from a high-dimensional input space to a usually two-dimensional output space while preserving topological relations as faithfully as possible. Input signals $x \in \Re^n$ are presented to the map, consisting of a grid of units with n-dimensional weight vectors, in random order. An activation function based on some metric (e.g. the Euclidean Distance) is used to determine the winning unit (the 'winner'). In the next step the weight vector of the winner as well as the weight vectors of the neighboring units are modified following some learning rate in order to represent the presented input signal more closely. Basically, the entries to be included in the library system are represented in the form of feature vectors, which are created by parsing the texts and processing the resulting word histograms to provide a compact and effective representation of the texts. These feature vectors are used as input vectors to train a standard SOM.

3 The Growing Hierarchical Self-Organizing Map

While the SOM has proven to be a very suitable tool for detecting structure in high-dimensional data and organizing it accordingly on a two-dimensional output space, some shortcomings have to be mentioned. These include its inability to capture the inherent hierarchical structure of data. Furthermore, the size of the map has to be determined in advance ignoring the characteristics of an (unknown) data distribution. These drawbacks have been addressed separately in several modified architectures of the SOM [1, 3, 8]. However, none of these approaches provides an architecture which fully adapts itself to the characteristics of the input data. To overcome the limitations of both fix-sized and non-hierarchically adaptive architectures we developed the GHSOM, which dynamically fits its multi-layered architecture according to the structure of the data.

The GHSOM has a hierarchical structure of multiple layers where each layer consists of several independent growing self-organizing maps. Starting from a top-level map, each map, similar to the Growing Grid model [3], grows in size in order to represent a collection of data at a certain level of detail. In particular, starting with an initial 2 × 2 SOM, rows and columns of units are added to those areas of the map where input discrimination is rather poor. After a certain improvement of the granularity of data representation is reached, the units are analyzed to see whether they represent the data at a specific minimum level of granularity. Those units that have too diverse input data mapped onto them are expanded to form a new small SOM at a subsequent layer, where the respective data shall be represented in more detail. The growth process of these new maps continues again in a Growing-Grid like fashion. Units representing an already rather homogeneous set of data, on the other hand, will not require any further expansion at subsequent layers. The resulting GHSOM thus is fully adaptive to reflect, by its very architecture, the hierarchical structure inherent in the data, allocating more space for the representation of inhomogeneous areas in the input space.

4 Data Set

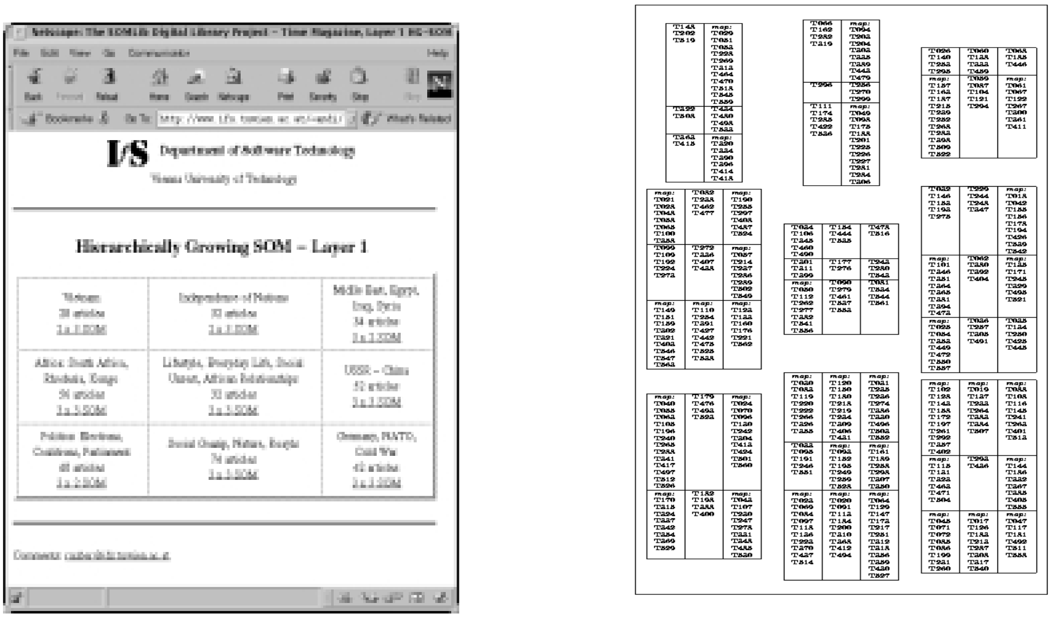
The SOMLib system and the GHSOM have been evaluated with numerous document collections, ranging from a small number of scientific abstracts to annual archives of newspapers comprising more than 40.000 articles, with documents in different languages. In the experiments presented hereafter we use the well-known *TIME Magazine* article collection as a reference document archive. The collection comprises 420 documents from the *TIME Magazine* of the early 1960's. The documents can be thought of as forming topical clusters in the high-dimensional feature space spanned by the words that the documents are made up of. The goal is to map and identify those clusters on the 2-dimensional map display. Thus, we use full-text indexing to represent the various documents according to the vector space model of information retrieval. The indexing process identified 5923 content terms, i.e. terms used for document representation, by omitting words that appear in more than 90% or less than 1% of the documents. The terms are roughly stemmed and weighted according to a $tf \times idf$, i.e. term frequency times inverse document frequency, weighting scheme [16], which assigns high values to terms that are considered important in describing the contents of a document. Following the feature extraction process we end up with 420 vectors describing the documents in the 5923-dimensional document space, which are further used for neural network training.

5 Experimental Results

Based on the single unit representing the mean of all data points at layer 0, the GHSOM training algorithm started with a 2 × 2 SOM at layer 1. The training process for this map continued with additional units being added until the quantization error fell below a certain percentage of the overall quantization error of the unit at layer 0. The resulting first-layer map is depicted in Figure 1a. The map has grown for two stages, adding one row and one column respectively, resulting in 3 × 3 units representing 9 major topics in the document collection.

For convenience we list the topics of the various units, rather then the individual articles in the figure. For example, we find unit (1/1) to represent all articles related to the situation in Vietnam, whereas Middle-East topics are covered on unit (1/3), or articles related to elections and other political topics on unit (3/1) in the lower left corner to name but a few.

Based on this first separation of the most dominant topical clusters in the article collection, further maps were automatically trained to represent the various topics in more detail. This results in 9 individual



Netscape: The SOMLib Digital Library Project – Time Magazine, Layer 1 GH-SOM
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Go To: http://www.ifs.tuwien.ac.at/~andi/ What's Related

IfS Department of Software Technology

Vienna University of Technology

Hierarchically Growing SOM – Layer 1

Vietnam 38 articles 2 x 3 SOM	Independence of Nations 32 articles 2 x 3 SOM	Middle East, Egypt, Iraq, Syria 34 articles 3 x 2 SOM
Africa: South Africa, Rhodesia, Congo 56 articles 3 x 3 SOM	Lifestyle, Everyday Life, Social Unrest, African Relationships 32 articles 3 x 3 SOM	USSR – China 52 articles 3 x 2 SOM
Politics: Elections, Coalitions, Parliament 48 articles 2 x 3 SOM	Social Change, Nations, Europe 74 articles 3 x 3 SOM	Germany, NATO, Cold War 42 articles 3 x 3 SOM

Comments: rauber@ifs.tuwien.ac.at

Figure 1: (a) Layer 1 and (b) Layer 2 of the *GH-SOM*

maps on layer 2, each representing the data of the respective higher-layer unit in more detail. Some of the units on these layer 2 maps were further expanded as distinct SOMs in layer 3.

The resulting layer 2 maps are depicted in Figure 1b. Please note, that – according to the structure of the data – the maps on the second layer have grown to different sizes, such as a small 2 × 2 map representing the articles of unit (3/1) of the first map, up to 3 × 3 maps for the units (2/1), (3/2) and (3/3). Taking a more detailed look at the first map of layer 2 representing unit (1/1) of layer 1, this map gives a clearer representation of articles covering the situation in Vietnam. Units (1/1) and (2/1) on this map represent articles on the fighting during the Vietnam War, whereas the remaining units represent articles on the internal conflict between the catholic government and buddhist monks. At this layer, the two units (1/2) and (3/2) have further been expanded to form separate maps with 3 × 3 units each at layer 3. These again represent articles on the war and the internal situation in Vietnam in more detail.

To give another example of the hierarchical structures identified during the *growing hierarchical SOM* training process, we may take a look at the 2×3 map representing the articles of unit (3/1) of the first layer map. All of these articles were found to deal with political matters on layer 1. This common topic is now displayed in more detail at the resulting second-layer map. For example, we find unit (1/3) to represent articles on the elections in India. Next to these, we find on units (1/2) and (2/3) articles covering the elections and discussions about political coalitions between Socialists and Christian Democrats in Italy. The remaining 3 units on this map deal with different issues related to the Profumo-Keeler scandal in Great Britain, covering the political hearings in parliament as well as background information on this scandal and the persons involved. Again, some of the units have been expanded at a further level of detail forming 3 × 2 or 3 × 3 SOMs on layer 3.

As a last example, consider the 3×3 map representing articles of unit (3/3) of the first layer. In the first layer we find this unit to cover articles related to east-west relationships, mainly dealing with post-war Germany, the relationships between Germany and the Soviet Union and the NATO. In the second-layer map we find these topics to be somewhat clearer separated, with units (1/1) to (1/3) covering mainly Germany-related articles, whereas the other two topics are represented by the remaining 4 units, most of which again are explained in more detail in the corresponding third-layer maps.

When comparing the GHSOM with a conventional flat counterpart we may identify the locations of the articles on the 9 second-layer maps on a corresponding 10 × 15 SOM [1]. This allows us to view the hierarchical structure of the data on the flat map.

[1] The map is available at `http://www.ifs.tuwien.ac.at/ifs/research/ir/` for interactive exploration.

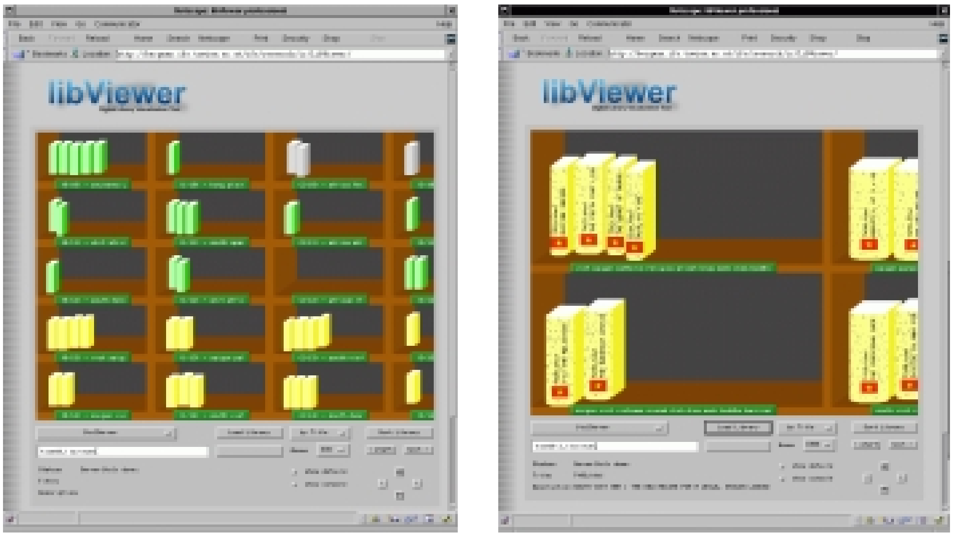


Figure 2: Detailed libViewer representation of *TIME* Magazine article collection SOM: lower left area (a) distant and (b) close-up representation

6 Other modules of the SOMLib System

Conventional approaches to document space representation using the SOM require all documents to be available at one place for network training. This prerequisite severly limits the applicability of these systems in real-world applications, where document collections are distributed across several sites or released at certain time intervals. If individual SOMs are trained on these collections, we would like to use the classification provided by these maps, and integrate those maps rather than train a new SOM, which would force us to collect all articles locally. Within the SOMLib system we thus devised a way to integrate different maps, or parts thereof, using their weight vectors rather than the actual document vectors for training [9, 15]. This apporach allows us to build a set of interconnected SOMs, using the classification provided at these sources.

While the SOM has found wide appreciation in the field of text classification, its application has been limited by the fact that the topics of the various cluster were not evident from the resulting mapping. In order to find out which topics are covered in certain areas of the map, the actual articles had to be read to find descriptive keywords for a cluster. To counter this problem, we developed the LabelSOM method, which analyses the trained SOM to automatically extract a set of attributes, i.e. keywords, that are most descriptive for a unit [7, 10]. Basically, the attributes showing a low quantization error value and a high weight vector value, comparable to a low variance and a high mean among all input vectors mapped onto a specific unit, are selected as labels. Thus, the various units are characterized by keywords describing the topics of the documents mapped onto them.

Last, but not least, while the spatial organization of documents on the 2-dimensional map in combination with the automatically extracted concept labels supports orientation in and understanding of an unknown document repository, much information on the documents cannot be told from the resulting representation. Information like the size of the underlying document, its type, the date it was created, when it was accessed for the last time and how often it has been accessed at all, its language etc. is not provided. We thus developed the libViewer, a metaphor-graphics based interface to a digital library [11]. Documents are thus no longer represented as textual listings, but as graphical objects of different *representation types* such as binder, papers, hardcover books, paperbacks etc, with further metadate information being conveyed by additional metaphors such as *spine width, logos, well-thumbed spines*, different degrees of *dustiness, highlighting glares, position in the shelf* and others.

An example for a libViewer library visualization is depicted in Figure 2, showing part of the Vietnam-Cluster of the flat SOM both inthe distant as well as the close-up view. Please note, that the relative age of a document, to pick just one example, become intuitively visible from the graphical representation, as does the relative size of a document, relieving the user from having to read and to interpret the metadata as textual descriptions.

7 Conclusions

In this paper we have argued in favour of a hierarchical representation of document archives. Such an organization provides a more intuitive means for exploring and understanding large information spaces. The *growing hierarchical self-organizing map* (GHSOM) has shown to provide this kind of representation by adapting both its hierarchical structure as well as the sizes of each individual map to represent data at desired levels of granularity. It fits its architecture according to the requirements of the input space, reliefing the user from having to define a static organization prior to the training proces. Multiple experiments have shown both its capabilities of hierarchically orginzing document collection according to their topics, as well as the benefits of providing a better overview of, especially, larger collections, where single map-based representations tend to become unacceptably large.

References

[1] J. Blackmore and R. Miikkulainen. Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map. In *Proc of the IEEE Int'l Conf on Neural Networks (ICNN'93)*, San Francisco, CA, USA, 1993.

[2] M. Dittenbach, D. Merkl, and A. Rauber. The growing hierarchical self-organizing map. In *Proc of the Int'l Joint Conf on Neural Networks (IJCNN 2000)*, Como, Italy, 2000.

[3] B. Fritzke. Growing Grid – A self-organizing network with constant neighborhood range and adaption strength. *Neural Processing Letters*, 2(5):1 – 5, 1995.

[4] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.

[5] T. Kohonen. *Self-organizing maps.* Springer-Verlag, Berlin, 1995.

[6] D. Merkl and A. Rauber. Alternative ways for cluster visualization in self-organizing maps. In *Proc of the Workshop on Self-Organizing Maps (WSOM97)*, Espoo, Finland, 1997.

[7] D. Merkl and A. Rauber. Automatic labeling of self-organizing maps for information retrieval. In *Proc of the 6. Int'l Conf on Neural Information Processing (ICONIP99)*, Perth, Australia, 1999.

[8] R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2:83 – 101, 1990.

[9] A. Rauber. SOMLib: A distributed digital library system based on self-organizing maps. In *Proc of the 10. Italian Workshop on Neural Nets (WIRN98)*, Springer Perspectives in Neural Computing, Vietri sul Mare, Italy, 1998.

[10] A. Rauber. LabelSOM: On the labeling of self-organizing maps. In *Proc of the Int'l Joint Conf on Neural Networks (IJCNN'99)*, Washington, DC, July 10. - 16. 1999.

[11] A. Rauber and H. Bina. Visualizing electronic document repositories: Drawing books and papers in a digital library. In *Advances in Visual Database Systems: Proc of the IFIP TC2 WG2.6 5. Working Conf on Visual Database Systems*, Fukuoka, Japan, May 2000.

[12] A. Rauber, M. Dittenbach, and D. Merkl. Automatically detecting and organizing documents into topic hierarchies: A neural-network based approach to bookshelf creation and arrangement. In *Proc of the 4. European Conf on Research and Advanced Technologies for Digital Libraries (ECDL2000)*, Springer LNCS 1923, Lisboa, Portugal, 2000.

[13] A. Rauber and D. Merkl. Automatic labeling of self-organizing maps: Making a treasure map reveal its secrets. In *Proc of the 3. Pacific-Asia Conf on Knowledge Discovery and Data Mining (PAKDD99)*, Springer LNCS/LNAI 1574 Beijing, China, 1999.

[14] A. Rauber and D. Merkl. The SOMLib Digital Library System. In *Proc of the 3. European Conf on Research and Advanced Technology for Digital Libraries (ECDL99)*, Springer LNCS 1696, Paris, France, 1999.

[15] A. Rauber and D. Merkl. Providing topically sorted access to subsequently released newspaper editions or: How to build your private digital library. In *Proc of the 11. Int'l Conf on Database and Expert Systems Applications (DEXA2000)*, Springer LNCS 1873, London, UK, 2000.

[16] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA, 1989.

[17] A. Ultsch. Self-organizing neural networks for visualization and classification. In *Information and Classification. Concepts, Methods and Application*, Dortmund, Germany, 1992.