CIA's view of the world and what neural networks learn from it:
A comparison of geographical document space representation metaphors

Dieter Merkl and Andreas Rauber

Institut für Softwaretechnik, Technische Universität Wien
Resselgasse 3/188, A–1040 Wien, Austria
{dieter, andi}@ifs.tuwien.ac.at

Abstract. Text collections may be regarded as an almost perfect application arena for unsupervised neural networks. This because many operations computers have to perform on text documents are classification tasks based on noisy patterns. 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 hierarchical feature maps to represent the contents of a document archive. These neural networks have a 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

Today's information age may be characterized by constant massive production and dissemination of written information. Powerful tools for exploring, searching, and organizing this mass of information are needed. Particularly the aspect of exploration has found only limited attention. Current information retrieval technology still relies on systems that retrieve documents based on the similarity between keyword-based document and query representations.

The map metaphor for displaying the contents of a document library in a two-dimensional map has gained some interest [3, 4, 5, 6] Maps are used to visualize the similarity between documents in terms of distances within the two-dimensional map display. Hence, similar documents may be found in neighboring regions of the map display.

The map metaphor, however, has its limitations in that each document is represented within a single two-dimensional map. Since the documents are described in a very high-dimensional feature space, the two-dimensional map representation has necessarily some imprecisions. 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.

A comparison with traditional document archives reveals that these archives are usually organized into hierarchies according to the subject matter of the various documents. This observation has stimulated research in the direction of using hierarchical clustering techniques based on statistical cluster analysis. The specific strengths and weaknesses of these approaches are well explored [9, 10].

In this paper we argue in favor of establishing a hierarchical organization of the document space based on an unsupervised neural network. More precisely, we show the effects of using the *hierarchical feature map* [8] for text archive organization. The distinguished feature of this model is its layered architecture where each layer consists of a number of independent *self-organizing maps* [1]. 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.

The remainder of this paper is organized as follows. In Section 2 we give a brief description of the architectures and the training rules of the neural networks used in this study. Section 3 is dedicated to a description of the text documents that constitute our experimental document library. Sections 4 and 5 provide the experimental results from document classification. The former describes the results from using the self-organizing map, i.e. library organization according to the map metaphor. The latter gives results from using the hierarchical feature map, i.e. library organization according to the atlas metaphor. Finally, in Section 6 we give some conclusions.

2 Topology preserving self-organizing networks

2.1 Self-organizing maps

The self-organizing map [1, 2] is one of the most prominent artificial neural network models adhering to the unsupervised learning paradigm. The model consists of a number of neural processing elements, i.e. units. Each of the units i

is assigned an n-dimensional weight vector m_i. It is important to note that the weight vectors have the same dimensionality as the input patterns.

The training process of self-organizing maps may be described in terms of input pattern presentation and weight vector adaptation. Each training iteration t starts with the random selection of one input pattern $x(t)$. This input pattern is presented to the self-organizing map and each unit determines its activation. Usually, the Euclidean distance between weight vector and input pattern is used to calculate a unit's activation. The unit with the lowest activation is referred to as the *winner*, c, of the training iteration, i.e. $m_c(t) = \min_i ||x(t) - m_i(t)||$. Finally, the weight vector of the *winner* as well as the weight vectors of selected units in the vicinity of the *winner* are adapted. This adaptation is implemented as a gradual reduction of the component-wise difference between input pattern and weight vector, i.e. $m_i(t+1) = m_i(t) \cdot \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)]$. Geometrically speaking, the weight vectors of the adapted units are moved a bit towards the input pattern. The amount of weight vector movement is guided by a so-called learning rate, α, decreasing in time. The number of units that are affected by adaptation is determined by a so-called neighborhood function, h_{ci}. This number of units also decreases in time.

This movement has as a consequence, that the Euclidean distance between those vectors decreases and thus, the weight vectors become more similar to the input pattern. The respective unit is more likely to win at future presentations of this input pattern. The consequence of adapting not only the *winner* alone but also a number of units in the neighborhood of the *winner* leads to a spatial clustering of similar input patters in neighboring parts of the self-organizing map. Thus, similarities between input patterns that are present in the n-dimensional input space are mirrored within the two-dimensional output space of the self-organizing map. The training process of the self-organizing map describes a topology preserving mapping from a high-dimensional input space onto a two-dimensional output space where patterns that are similar in terms of the input space are mapped to geographically close locations in the output space.

Consider Figure 1 for a graphical representation of self-organizing maps. The map consists of a square arrangement of 7×7 neural processing elements, i.e. units, shown as circles on the left hand side of the figure. The black circle indicates the unit that was selected as the *winner* for the presentation of input pattern $x(t)$. The weight vector of the *winner*, $m_c(t)$, is moved towards the input pattern and thus, $m_c(t+1)$ is nearer to $x(t)$ than was $m_c(t)$. Similar, yet less strong, adaptation is performed with a number of units in the vicinity of the *winner*. These units are marked as shaded circles in Figure 1. The degree of shading corresponds to the strength of adaptation. Thus, the weight vectors of units shown with a darker shading are moved closer to x than units shown with a lighter shading.

2.2 Hierarchical feature maps

The key idea of hierarchical feature maps as proposed in [8] is to use a hierarchical setup of multiple layers where each layer consists of a number of independent

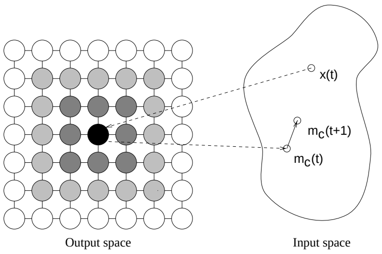

Fig. 1. Architecture of a 7×7 self-organizing map

self-organizing maps. One self-organizing map is used at the first layer of the hierarchy. For every unit in this map a self-organizing map is added to the next layer of the hierarchy. This principle is repeated with the third and any further layers of the hierarchical feature map. In Figure 2 we provide an example of a hierarchical feature map with three layers. The first layer map consists of 2×2 units, thus we find four independent self-organizing maps on the second layer. Since each map on the second layer consists again of 2×2 units, there are 16 maps on the third layer.

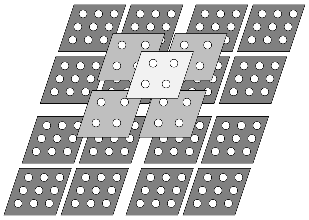

Fig. 2. Architecture of a three-layer hierarchical feature map

The training process of hierarchical feature maps starts with the self-organizing map on the first layer. This map is trained according to the standard training process of self-organizing maps as described above. When this first self-organizing map is stable, i.e. only minor further adaptation of the weight vectors are recorded, training proceeds with the maps of the second layer. Here, each map is trained with only that portion of the input data that is mapped on the respective unit in the higher layer map. By this, the amount of training data for a particular self-organizing map is reduced on the way down the hierarchy. Additionally, the vectors representing the input patterns may be shortened on the transition from one layer to the next. This shortage is due to the fact that some input vector components can be expected to be equal among those input data that are mapped onto the same unit. These equal components may be omitted for training the next layer maps without loss of information. There is no loss of information because these portions of the input vector are already represented by the higher layer unit.

2.3 Comparison of both models

Hierarchical feature maps have two benefits over self-organizing maps which make this model particularly attractive in an information retrieval setting as described in the remainder of this paper.

First, hierarchical feature maps have substantially shorter training times than self-organizing maps. The reason for that is twofold. On the one hand, there is the obvious input vector dimension reduction on the transition of one layer to the next. Shorter input vectors lead directly to reduced training times because of faster *winner* selection and weight vector adaptation. On the other hand, and even more important, the self-organizing training process is performed faster because the spatial relation of different areas of the input space is maintained by means of the network architecture rather than by means of the training process. A more detailed treatment of this issue may be found in [7].

Second, hierarchical feature maps may be used to produce isolated, i.e. disjoint, clusters of the input data. Moreover, these disjoint clusters are gradually refined when moving down along the hierarchy. Contrary to that, in its basic form, the self-organizing map cannot be used to produce isolated clusters. The separation of data items rather is a tricky task that requires some insight into the structure of the input data. What one gets, however, from a self-organizing map is a sort of overall representation of input data similarities. In this sense we may use the following picture to contrast the two models of neural networks. Self-organizing maps can be used to produce maps of the input data whereas hierarchical feature maps produce an atlas of the input data. Taking up this metaphor the difference between both models is quite obvious. Self-organizing maps, in our point of view, provide the user with a single picture of the underlying data archive. As long as the map is not too large, this picture may be sufficient. As the maps grow larger, however, they have the tendency of providing too little orientation for the user. In such a case we would advise to change to hierarchical feature maps as the model for representing the contents of the data archive. In this case, the data is organized hierarchically which facilitates browsing into relevant portions of the data archive. In much the same way as one would probably not use the map of the world in order to find one's way from *Schönbrunn* to *Neustift* one would probably not use a single map of a document archive to find a particular document. Conversely, when given an atlas one might follow the hierarchy of maps along a path such as *World* → *Europe* → *Austria* → *Vienna* in order to finally find the way from *Schönbrunn* to *Neustift*. In a similar way an atlas of a document archive might be used.

3 Document representation

For the experiments presented thereafter we used the 1990 edition of the *CIA World Factbook* (http://www.odci.gov/cia/publications/factbook) as a sample document archive. The *CIA World Factbook* represents a text collection containing information on countries and regions of the world. The information is split into different categories such as *Geography*, *People*, *Government*, *Economy*, *Communications*, and *Defense Forces*.

We used full-text indexing to represent the various documents. The complete information on each country is used for indexing. In other words, for the present set of experiments we refrained from identifying the various document segments that contain the information on the various categories. In total, the 1990 edition of the *CIA World Factbook* consists of 245 documents. The indexing process identified 959 content terms, i.e. terms used for document representation. During indexing we omitted terms that appear in less than 15 documents or more than 196 documents. These terms are weighted according to a simple $tf \times idf$ weighting scheme [9]. With this indexing vocabulary the documents are represented according to the vector-space model of information retrieval. The various vectors representing the documents are further used for neural network training.

4 A map of the world

Based on the document description as outlined above, we first trained a 10×10 self-organizing map to represent the contents of the document archive. Figure 3 gives a graphical representation of the training result. For ease of identifying the various rows of units in the graphical representation, we separated these rows by horizontal lines. Each unit is either marked by a number of countries (or regions) or by a dot. The name of a country appears if this unit serves as the *winner* for that particular country (or more precisely for the input vector representing that country). A dot appears if the unit is never selected as *winner* for any document.

Figure 3 shows that the self-organizing map was quite successful in arranging the various input data according to their mutual similarity. It should be obvious that in general countries belonging to similar geographical regions are rather similar with respect to the different categories described in the *CIA World Factbook*. These geographical regions can be found in the two-dimensional map display as well. In order to ease the interpretation of the self-organizing map's training result, we have marked several regions manually. For example, the area on the left hand side of the map is allocated for documents describing various islands. We should note, that the *CIA World Factbook* contains a large number of descriptions of islands. It is interesting to note, that the description of the oceans can be found in a map region neighboring the area of islands.

In the lower center of the map we find the European countries. The cluster representing these countries is further decomposed into a cluster of small countries, e.g. *San Marino* and *Liechtenstein*, a cluster of Western European countries, and finally a cluster of Eastern European countries. The latter cluster

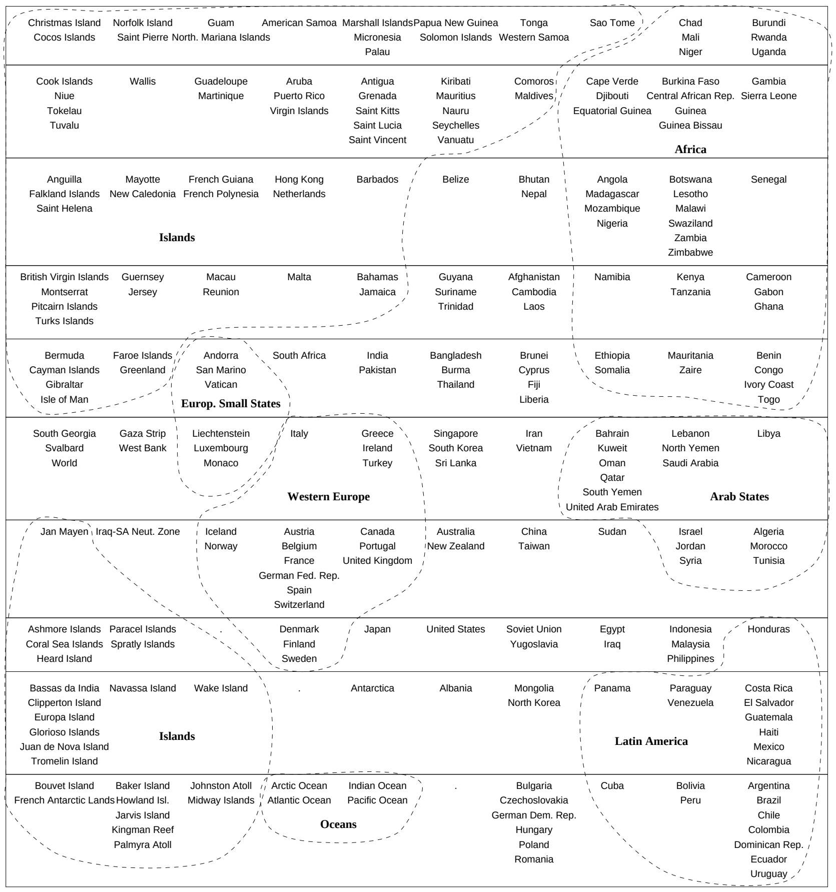

Fig. 3. 10×10 map of the world

is represented by a single unit in the last row of the output space. This unit has as neighbors other countries that are usually attributed as belonging to the Communist hemisphere, e.g. *Cuba*, *North Korea*, *Albania*, and *Soviet Union*. At this point it is important to recall that our document archive is the 1990 edition of the *CIA World Factbook*. Thus, the descriptions refer to a time before the "fall" of the Communist hemisphere.

Other clusters of interest are the region containing countries from Latin America (lower right of the map), the cluster containing Arab countries (middle right of the map), or the cluster of African countries (upper right of the map).

Overall, the representation of the document space is highly successful in that similar documents are located close to one another. Thus, it is easy to find an orientation in this document space. The negative point, however, is that each document is represented on the very same map. Since the self-organizing map represents a very high-dimensional data space (959 index terms) within a two-dimensional display it is only natural that some information gets lost during the mapping process. As a consequence, it is rather difficult to identify the various clusters. Imagine Figure 3 without the dashed lines indicating cluster boundaries. Without this information it is only possible to identify, say, African countries when prior information about the document collection is available.

5 An atlas of the world

In the previous section we have described the results from using self-organizing maps with the data of the *CIA World Factbook*. The major shortcoming of this neural network model is that the various documents are represented within only one two-dimensional output space making it difficult to identify cluster boundaries without profound insight into the underlying document collection.

The hierarchical feature map can provide essential assistance in isolating the different clusters. The isolation of clusters is achieved thanks to the architecture of the neural network which consists of layers of independent self-organizing maps. Thus, in the highest layer the complete document archive is represented by means of a small map (in terms of the number of neural processing elements). Each unit is then further developed within its own branch of the neural network.

For the experiment presented thereafter we used a setup of the hierarchical feature map using four layers. The respective maps have the following dimensions: 3×3 on the first layer, 4×4 on the second layer, and 3×3 on the third and fourth layer. This setup has been determined empirically.

Figure 4 presents the contents of the first layer self-organizing map. In order to keep the information at a minimum we refrained from typing the names of the various countries in this figure. We rather present some aggregated information concerning the geographical location of the various countries.

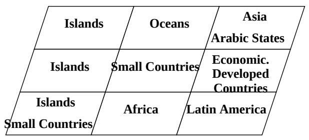

Fig. 4. Hierarchical feature map: First layer

In the remainder of this discussion we will just present the branch of the hierarchical feature map that contains what we called *economically developed countries*. The other branches cannot be shown in this paper because of space considerations. These branches, however, are formed quite similarly.

In Figure 5 we show the arrangement of the second layer within the branch of *economically developed countries*. In this map, the various countries are separated roughly according to either their geographic location or their political system. The clusters are symbolized by using different shades of grey.

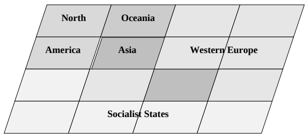

Fig. 5. Hierarchical feature map: Second layer

Finally, Figure 6 shows the full-blown branch of *economically developed countries*. In this case it is straight-forward to identify the various cluster boundaries in that each cluster is represented by an individual self-organizing map. Higher level similarities are shown in higher levels of the hierarchical feature map.

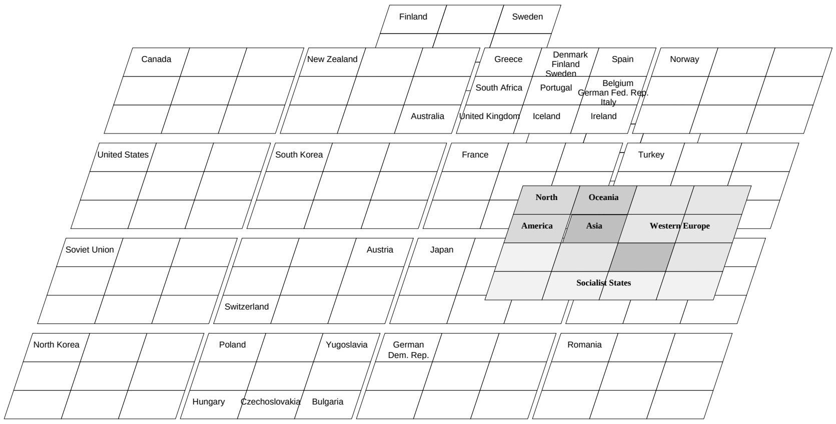

Fig. 6. A subset of the world map: Economically developed countries

6 Conclusions

In this paper we have provided an account on the feasibility of using unsupervised neural networks in a highly important task of information retrieval, namely text classification. As an experimental document collection we used the description of various countries as contained in the 1990 edition of the *CIA World Factbook*. For this document collection it is rather easy to judge the quality of the classification result. For document representation we relied on the vector space model and a simple $tf \times idf$ term weighting scheme.

We demonstrated that both the map and the atlas metaphor are highly useful for assisting the user to find her orientation within the document space. The shortcoming of the map metaphor, however, is that each document is shown in one large map and thus, the borderline between clusters of related and clusters of unrelated documents are sometimes hard to find. This is especially the case if the user does not have sufficient insight into the contents of the document collection.

The atlas metaphor overcomes this limitation in that the clusters of documents are clearly visible because of the architecture of the neural network. The document space is separated into independent maps along different layers in a hierarchy. The user thus gets the best of both worlds. The similarity between documents is shown in a fine-grained level in maps of the lower layers of the hierarchy while the overall organizational principles of the document archive are shown at higher layer maps. Since such a hierarchical arrangement of documents is the common way of organizing conventional libraries, only small intellectual overhead is required from the user to find her way through the document space.

References

1. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.
2. T. Kohonen. *Self-organizing maps.* Springer-Verlag, Berlin, 1995.
3. T. Kohonen, S. Kaski, K. Lagus, and T. Honkela. Very large two-level SOM for the browsing of newsgroups. In *Proc Int'l Conf on Artificial Neural Networks (ICANN'96)*, Bochum, Germany, 1996.
4. K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In *Proc Int'l Conf on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, 1996.
5. X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proc Int'l ACM SIGIR Conf on R&D in Information Retrieval (SIGIR'91)*, Chicago, IL, 1991.
6. D. Merkl. A connectionist view on document classification. In *Proc Australasian Database Conf (ADC'95)*, Adelaide, SA, 1995.
7. D. Merkl. Exploration of text collections with hierarchical feature maps. In *Proc Int'l ACM SIGIR Conf on R&D in Information Retrieval (SIGIR'97)*, Philadelphia, PA, 1997.
8. R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2, 1990.
9. G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA, 1989.
10. P. Willet. Recend trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24, 1988.

This article was processed using the LaTeX macro package with LLNCS style.