Document Classification with Unsupervised Artificial Neural Networks

Dieter Merkl and Andreas Rauber

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

Abstract. Text collections may be regarded as an almost perfect application arena for unsupervised neural networks. This is 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.

An attractive way to assist the user in document archive exploration is based on unsupervised artificial neural networks for document space representation. A number of research publications show that this idea has found appreciation in the community [23–27,31,33,44]. 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.

This map metaphor for document space visualization, however, has its limitations in that each document is represented within one single two-dimensional map. Since the documents are described in a very high-dimensional feature space constituted by the index terms representing the contents of the documents, 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 of the document space 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 his or her way along the library choosing 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 information retrieval in the direction of using hierarchical clustering techniques based on statistical cluster analysis. The specific strengths and weaknesses of these approaches are well explored [45,49]. An interesting recent approach is Scatter/Gather that relies on clustering during query processing [12].

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* [36] 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* [21]. 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. Maps from lower layers of the hierarchy are used to provide fine-grained distinction between individual documents. Such an organization 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 work 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. In Section 6 we give a brief review of related research on document classification with self-organizing maps and other artificial neural network models adhering to the unsupervised learning paradigm. Finally, in Section 7 we present some conclusions.

2 Topology preserving self-organizing neural networks

2.1 Self-organizing maps

The self-organizing map [21,22] 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, $m_i \in \Re^n$. 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 the weight vector and the input pattern is used to calculate a unit's activation. In this particular case, the unit with the lowest activation is referred to as the *winner*, c, of the training iteration, as given in Expression (1).

$$c : m_c(t) = \min_i ||x(t) - m_i(t)|| \qquad (1)$$

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 difference between corresponding components of the input pattern and the weight vector, as shown in Expression (2).

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \qquad (2)$$

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 such that towards the end of the training process only the *winner* is adapted. Typically, the neighborhood function is a unimodal function which is symmetric around the location of the winner and monotonically decreasing with increasing distance from the winner. A Gaussian may be used to model the neighborhood function as given in Expression (3) with r_i representing the two-dimensional vector pointing to the location of unit i within the grid, and $||r_c - r_i||$ denoting the distance between units c, i.e. the *winner* of the current training iteration, and i in terms of the output space. It is common practice that at the beginning of training a wide area of the output space is subject to adaptation. The spatial width of units affected by adaptation is reduced gradually during the training process. Such a strategy allows the formation of large clusters at the beginning and fine-grained input discrimination towards the end of the training process. The spatial width of adaptation is guided by means of the time-varying parameter σ.

$$h_{ci}(t) = \exp\left(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}\right) \tag{3}$$

The movement of weight vectors has the consequence, that the Euclidean distance between input and weight 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 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 $x(t)$ than units shown with a lighter shading.

2.2 Hierarchical feature maps

The key idea of hierarchical feature maps as proposed in [36,38] is to use a hierarchical setup of multiple layers where each layer consists of a number of independent 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

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

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 observed, 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 (almost) 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 the omitted 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 from 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 [32].

Second, hierarchical feature maps may be used to produce disjoint clusters of the input data. Moreover, these disjoint clusters are gradually refined when moving down along the hierarchy. Contrary to that, the self-organizing map in its basic form cannot be used to produce disjoint clusters. The separation of data items is a rather tricky task that requires some insight into the structure of the input data.

What one gets, however, from a self-organizing map is an 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.

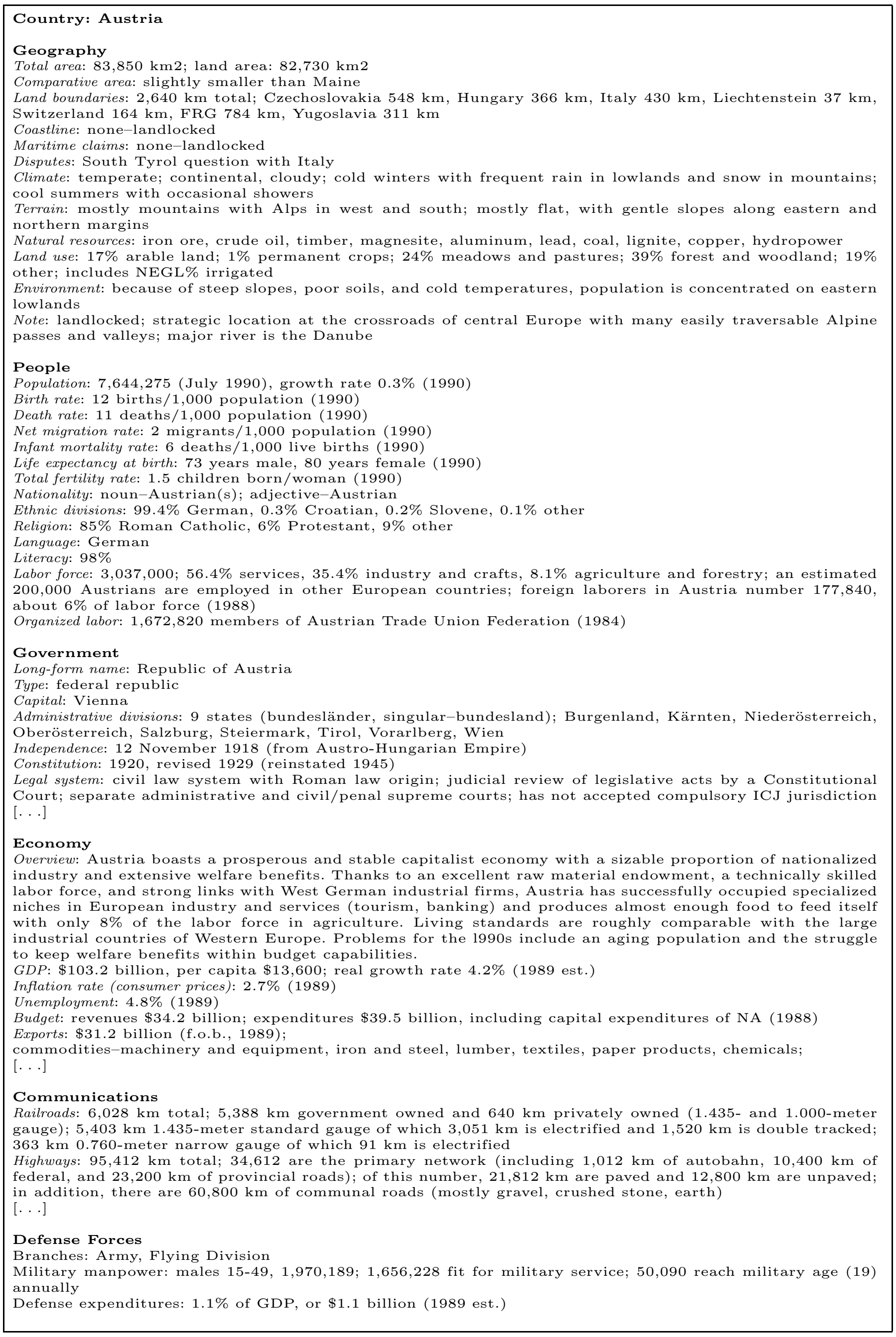
3 The experimental document collection

Generally, the task of text classification aims at uncovering the semantic similarities between various documents. In a first step, the documents have to be mapped onto some representation language in order to enable further analyses. This process is termed indexing in the information retrieval literature. A number of different strategies have been suggested over the years of information retrieval research. Still one of the most common representation techniques is single term full-text indexing where the text of the documents is accessed and the various words forming the document are extracted. These words may be mapped to their (often just approximate) word stem yielding the so-called terms used to represent the documents. The resulting set of terms is usually cleared from so-called stop-words, i.e. words that appear either too often or too rarely within the document collection and thus have only little influence on discriminating between different documents and would just unnecessarily increase the computational load during classification.

In a vector-space model of information retrieval the documents contained in a collection are represented by means of feature vectors x of the form $x = [\xi_1, \xi_2, \ldots, \xi_n]^T$. In such a representation, the ξ_i, $1 \leq i \leq n$, correspond to the index terms extracted from the documents as described above. The specific value of ξ_i corresponds to the importance of index term i in describing the particular document at hand. One might find a lot of strategies to prescribe the importance of an index term for a particular document [46]. Without loss of generality, we may assume that this importance is represented as a scalar in the range of $[0, 1]$ where *zero* means that this particular index term is absolutely unimportant to describe the document. Any deviation from *zero* towards *one* is proportional to the increased importance of the index term at hand. In such a vector-space model, the similarity between two text documents corresponds to the distance between their vector representations [47].

For the experiments presented thereafter we use 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*. Consider as an example the description of *Austria* as given in Figure 3.

We use 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



Country: Austria

Geography
Total area: 83,850 km2; land area: 82,730 km2
Comparative area: slightly smaller than Maine
Land boundaries: 2,640 km total; Czechoslovakia 548 km, Hungary 366 km, Italy 430 km, Liechtenstein 37 km, Switzerland 164 km, FRG 784 km, Yugoslavia 311 km
Coastline: none–landlocked
Maritime claims: none–landlocked
Disputes: South Tyrol question with Italy
Climate: temperate; continental, cloudy; cold winters with frequent rain in lowlands and snow in mountains; cool summers with occasional showers
Terrain: mostly mountains with Alps in west and south; mostly flat, with gentle slopes along eastern and northern margins
Natural resources: iron ore, crude oil, timber, magnesite, aluminum, lead, coal, lignite, copper, hydropower
Land use: 17% arable land; 1% permanent crops; 24% meadows and pastures; 39% forest and woodland; 19% other; includes NEGL% irrigated
Environment: because of steep slopes, poor soils, and cold temperatures, population is concentrated on eastern lowlands
Note: landlocked; strategic location at the crossroads of central Europe with many easily traversable Alpine passes and valleys; major river is the Danube

People
Population: 7,644,275 (July 1990), growth rate 0.3% (1990)
Birth rate: 12 births/1,000 population (1990)
Death rate: 11 deaths/1,000 population (1990)
Net migration rate: 2 migrants/1,000 population (1990)
Infant mortality rate: 6 deaths/1,000 live births (1990)
Life expectancy at birth: 73 years male, 80 years female (1990)
Total fertility rate: 1.5 children born/woman (1990)
Nationality: noun–Austrian(s); adjective–Austrian
Ethnic divisions: 99.4% German, 0.3% Croatian, 0.2% Slovene, 0.1% other
Religion: 85% Roman Catholic, 6% Protestant, 9% other
Language: German
Literacy: 98%
Labor force: 3,037,000; 56.4% services, 35.4% industry and crafts, 8.1% agriculture and forestry; an estimated 200,000 Austrians are employed in other European countries; foreign laborers in Austria number 177,840, about 6% of labor force (1988)
Organized labor: 1,672,820 members of Austrian Trade Union Federation (1984)

Government
Long-form name: Republic of Austria
Type: federal republic
Capital: Vienna
Administrative divisions: 9 states (bundesländer, singular–bundesland); Burgenland, Kärnten, Niederösterreich, Oberösterreich, Salzburg, Steiermark, Tirol, Vorarlberg, Wien
Independence: 12 November 1918 (from Austro-Hungarian Empire)
Constitution: 1920, revised 1929 (reinstated 1945)
Legal system: civil law system with Roman law origin; judicial review of legislative acts by a Constitutional Court; separate administrative and civil/penal supreme courts; has not accepted compulsory ICJ jurisdiction
[. . .]

Economy
Overview: Austria boasts a prosperous and stable capitalist economy with a sizable proportion of nationalized industry and extensive welfare benefits. Thanks to an excellent raw material endowment, a technically skilled labor force, and strong links with West German industrial firms, Austria has successfully occupied specialized niches in European industry and services (tourism, banking) and produces almost enough food to feed itself with only 8% of the labor force in agriculture. Living standards are roughly comparable with the large industrial countries of Western Europe. Problems for the l990s include an aging population and the struggle to keep welfare benefits within budget capabilities.
GDP: $103.2 billion, per capita $13,600; real growth rate 4.2% (1989 est.)
Inflation rate (consumer prices): 2.7% (1989)
Unemployment: 4.8% (1989)
Budget: revenues $34.2 billion; expenditures $39.5 billion, including capital expenditures of NA (1988)
Exports: $31.2 billion (f.o.b., 1989);
commodities–machinery and equipment, iron and steel, lumber, textiles, paper products, chemicals;
[. . .]

Communications
Railroads: 6,028 km total; 5,388 km government owned and 640 km privately owned (1.435- and 1.000-meter gauge); 5,403 km 1.435-meter standard gauge of which 3,051 km is electrified and 1,520 km is double tracked; 363 km 0.760-meter narrow gauge of which 91 km is electrified
Highways: 95,412 km total; 34,612 are the primary network (including 1,012 km of autobahn, 10,400 km of federal, and 23,200 km of provincial roads); of this number, 21,812 km are paved and 12,800 km are unpaved; in addition, there are 60,800 km of communal roads (mostly gravel, crushed stone, earth)
[. . .]

Defense Forces
Branches: Army, Flying Division
Military manpower: males 15-49, 1,970,189; 1,656,228 fit for military service; 50,090 reach military age (19) annually
Defense expenditures: 1.1% of GDP, or $1.1 billion (1989 est.)

Fig. 3. CIA World Factbook: Country description of Austria

according to a simple $tf \times idf$ weighting scheme [45], i.e. term frequency times inverse document frequency. 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 4 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). Contrary to that, a dot appears if the unit is never selected as *winner* for any document.

Figure 4 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 see, that the description of the oceans can be found in a map region neighboring the area of islands in the lower middle part of the map.

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 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).

Fig. 4. 10 × 10 map of the world

For the sake of honesty, however, we have to note that the result from document classification with self-organizing maps has some imprecisions. Consider the above mentioned cluster of Western European countries. Canada, contained in this cluster, is certainly misclassified with respect to its geographical location. Yet, its economic situation might have been the reason for this specific placement. We have no intuitive explanation, however, why The Netherlands are mapped right within the Islands cluster in the upper left part of the map. This country is placed on the very same unit as is Hong Kong.

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 4 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 hereafter 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 after a series of training runs. It is certainly a shortcoming of this particular artificial neural network model that the architecture has to be defined before training begins. In order to do this, one has to have some understanding of the underlying data material. However, we are currently addressing this issue in that we are working towards an incrementally growing version of the hierarchical feature map where the architecture will be defined as a result of the unsupervised learning process such that no prior information will be needed. In particular, the depth of the hierarchy as well as the size of the various layers will be determined during training. Our first experience with this new artificial neural network model is highly encouraging in that similar results as those presented in this work are obtained.

Figure 5 presents the contents of the first layer self-organizing map. In order to keep the information at a minimum we refrained from showing the names of the various countries in this figure. We rather present some aggregated information concerning the various countries.

Fig. 5. 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 work because of space considerations. These branches, however, are formed quite similarly.

In Figure 6 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 geographical location or their political system. The clusters are symbolized by using different shades of grey.

Fig. 6. Hierarchical feature map: Second layer

Finally, Figure 7 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.

6 Related work

Document classification by using artificial neural networks has already gained some attention in the information retrieval community. Among the first and most influential papers we certainly have to mention [2,3]. In this work the author argues in favor of using feed-forward neural networks for query expansion. The neural network's role within the overall system is to perform

Fig. 7. A subset of the world map: Economically developed countries

spreading-activation during retrieval in order to describe the relation between terms, on the one hand, and documents and queries, on the other hand. This line of research is continued in [42,43]. Comparable work is described in [19,48]. Another approach relying on feed-forward networks is reported in [9]. In this paper the author describes an experiment where the weights of the neural networks are computed by using a supervised learning strategy rather than set directly using the term frequency histograms as it is done in most of the other studies.

A different line of research is performed using unsupervised neural networks. The paper of Lin et al. [26] perhaps marks the first attempt to utilize unsupervised neural networks for an information retrieval task. Similar to our approach, the authors rely on self-organizing maps. In this paper, however, the document representation is made up from 25 manually selected index terms and is thus not really realistic. In [27] this line of research is continued, yet this time with full-text indexed documents.

Among the shortcomings of self-organizing maps one certainly has to mention the remarkable computational demands of the learning rule. Possibilities to increase the speed of learning may be found in the learning rule itself by using the biologically motivated concept of lateral inhibition [18]. Two different realizations of this principle are described in [30,37]. Pragmatically speaking, a learning function incorporating lateral inhibition pushes the weight vector of units distant from the winner slightly away from the current input pattern. The effect of such a learning function is that the phase of rough input clustering is considerably accelerated in terms of the number of learning iterations that are needed to reach a stable state of the self-organizing process. Another convenient behaviour of such a learning rule is a remarkably increased accuracy of pattern representation in terms of the remaining quantization error after completion of the training process.

Another means for increasing the speed of the learning process is obviously related to the representation of the documents. In general, the feature space is not free from correlations due to the inexact mapping of free-form natural language text onto lexical entities. As a consequence, one might be interested in the transformation of the original document representation into a (much) lower dimensional space. In fact, this is the underlying principle of the *latent semantic indexing* technique [10]. Comparable results might be achieved by using *principal component analysis* [17] in order to reduce the dimensionality of the feature space. We refer to [1] for a recent report on using principal component analysis in the area of document processing. An approximation to the principal components may be gained by utilizing auto-associative feed-forward neural networks, i.e. feed-forward networks trained to reproduce the input at their output layer via a smaller layer of hidden units. The smaller hidden layer is further used to represent the input patterns. The effect of such a dimension reduction in keyword-based document representation and subsequent self-organizing map training with the compressed input patterns is described in [28]. To summarize the results, the experiments indicated that basically the same cluster results can be achieved by spending only a fraction of time for the training process.

Only recently, a number of papers have been published on the utilization of the self-organizing map for text representation [16] based on the seminal work of [41] and subsequent interactive exploration [14,15,24,25], i.e. the *WEBSOM* project. One of the interesting aspects of this project is the radically different document representation methodology. Pragmatically speaking, the co-occurrence of words in a document is analyzed by means of the self-organizing map leading to a word category map which is further used to represent the various documents contained in the text archive. The generation of these word category maps is described in [13,16]. Comparable work on word category maps, yet relying on a more conventional co-occurrence representation, is reported in [35].

In our most recent work we were particularly interested in two issues. First, we investigated the feasibility of self-organizing maps for organizing distributed document archives [39]. In this paper we argue in favor of representing the various portions of the document archive by means of self-organizing maps that may be integrated in order to give an overview of the complete archive. Second, we developed a method for automatically assigning labels to the units of the self-organizing map [34,40]. The labels are derived by analyzing the term co-occurrence patterns within documents mapped onto the same neural unit. These labels give very clear hints on the contents of the documents and thus facilitate the interpretation of training results.

Apart from the self-organizing map just a limited number of other unsupervised models have been evaluated for their usability in information retrieval applications. In [20,50] the authors report on an application of *growing cell structures*, a network with adaptive architecture [11]. The learning process of this artificial neural network is highly similar to self-organizing maps in that during each training cycle the weight vector of the winner and those of a number of units in the neighborhood of the winner are adapted. A slight variation concerns the definition of the neighborhood where only direct neighbors of the winner are taken into account. The fundamental difference, however, is that ever after a fixed number of training cycles a new unit is added to the network at the position of the highest quantization error, i.e. the position of the largest deviation between the weight vector and the input patterns that are represented by that very unit. Additionally, a unit that serves the least often as winner may be deleted from the network. As an effect of this learning strategy the network structure itself is adapted to the particular requirements of the input space as opposed to self-organizing maps where the network structure in terms of the number of units and their topology has to be defined prior to the training process. This model, however, requires much more learning parameters, related to network structure adaptation, to be adjusted in advance. Moreover, it is much more susceptible to minor variations in these parameters than the self-organizing map.

A report on the applicability of *adaptive resonance theory* networks [6] to document clustering is provided in [29]. The major advantage of this type of network is its fast learning speed combined with continuous plasticity, i.e. the network is capable to add new data items without the need of re-training. In its most rudimentary form an adaptive resonance theory network consists of two layers, the one representing the input pattern and the other representing the various clusters in terms of a number of competitive units. The distinguished characteristic of that type of artificial neural networks, i.e. the continuous plasticity, is achieved by adding a new competitive unit in case none of the existing ones represents the actual input pattern with satisfying accuracy. In this sense, we might regard adaptive resonance theory networks as one of the earliest artificial neural network models with both adaptive weights and adaptive architecture. Information concerning intercluster similarity, however, cannot be deduced from the results.

There are still a number of apparently usable artificial neural network models unexplored as far as their applicability to document clustering is concerned. In particular, the so-called *generative topographic mapping* as only recently suggested in [4,5] is developed as a substitute to the widely used self-organizing maps. Basically, this neural network describes a latent variable density model with a sound statistical foundation which is claimed to have several advantageous properties when compared to self-organizing maps, but no significant disadvantages. Probably one of the more important advantages is that generative topographic mapping should be open for rigorous mathematical treatment, an area where the self-organizing map has a remarkable tradition in effective resistance [7,8]. We can imagine that an alternative model that lends itself to thorough mathematical treatment might reduce the highly time-consuming need for large numbers of empirical tests in order to realize a successful artificial neural network application. As just one example consider the question of how many neural units shall be used to represent a particular set of input patterns with satisfactory accuracy. An (approximate) answer may only be found empirically with self-organizing maps and hierarchical feature maps.

On balance, unsupervised neural networks have proven to be remarkably successful as tools for explorative analysis of document archives as a number of studies have demonstrated that unsupervised neural networks are highly capable in uncovering similarities between text documents.

7 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 self-organizing map and the hierarchical feature map are highly useful for assisting the user to find his or her orientation within the document space. The shortcoming of the self-organizing map, 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 hierarchical feature map 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 his or her way through the document space.

Acknowledgments

We like to thank the anonymous reviewer for his or her insightful comments on an earlier version of this work.

References

1. T. Bayer, I. Renz, M. Stein, and U. Kressel. Domain and language independent feature extraction for statistical text categorization. In *Proc of the Workshop on Language Engineering for Document Analysis and Recognition*, Sussex, United Kingdom, 1996.
2. R. K. Belew. A connectionist approach to conceptual information retrieval. In *Proc of the Int'l Conference on Artificial Intelligence and Law (ICAIL'87)*, Boston, MA, 1987.
3. R. K. Belew. Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In *Proc of the ACM SIGIR Int'l Conf on Research and Development in Information Retrieval (SIGIR'89)*, Cambridge, MA, 1989.
4. C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: A principled alternative to the self-organizing map. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'96)*, Bochum, Germany, 1996.
5. C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. Technical Report NCRG/96/015, Aston University, Neural Computing Research Group, http://www.ncrg.aston.ac.uk, Birmingham, United Kingdom, 1996.
6. G. A. Carpenter and S. Grossberg. The ART of adaptive pattern recognition by a self-organizing neural network. *IEEE Computer*, 21(3), 1988.
7. M. Cottrell and J.-C. Fort. Etude d'un processus d'auto-organisation. *Annales de l'Institut Henri Poincaré*, 23(1), 1987.
8. M. Cottrell, J.-C. Fort, and G. Pagès. Two or three things that we know about the Kohonen algorithm. In *Proc of the European Symposium on Artificial Neural Networks (ESANN'94)*, Bruxelles, Belgium, 1994.
9. F. Crestani. Learning strategies for an adaptive information retrieval system using neural networks. In *Proc of the IEEE Int'l Conf on Neural Networks (ICNN'93)*, San Francisco, California, 1993.
10. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Hashman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 1990.
11. B. Fritzke. Growing Cell Structures: A self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9), 1994.
12. M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proc Int'l ACM SIGIR Conf on R&D in Information Retrieval (SIGIR'96)*, Zurich, Switzerland, 1996.
13. T. Honkela. Self-organizing maps of words for natural language processing applications. In *Proceedings International ICSC Symposium on Soft Computing*, Nimes, France, 1997.
14. T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.
15. T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. WEBSOM — self-organizing maps of document collections. In *Proceedings Workshop on Self-Organizing Maps*, Espoo, Finland, 1997.

16. T. Honkela, V. Pulkki, and T. Kohonen. Contextual relations of words in Grimm tales analyzed by self-organizing maps. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'95)*, Paris, France, 1995.
17. I. T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, Berlin, 1986.
18. E. R. Kandel, S. A. Siegelbaum, and J. H. Schwartz. Synaptic transmission. In E. R. Kandel, J. H. Schwartz, and T. M. Jessell, editors, *Principles of Neural Science*. Elsevier, New York, 1991.
19. S. Keane, V. Ratnaike, and R. Wilkinson. Hierarchical news filtering. In *Proc of the Int'l Conf on Practical Aspects of Knowledge Management*, Basel, Switzerland, 1996.
20. M. Köhle and D. Merkl. Visualizing similarities in high dimensional input spaces with a growing and splitting neural network. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'96)*, Bochum, Germany, 1996.
21. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.
22. T. Kohonen. *Self-organizing maps.* Springer-Verlag, Berlin, 1995.
23. T. Kohonen. Self-organization of very large document collections: State of the art. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.
24. T. Kohonen, S. Kaski, K. Lagus, and T. Honkela. Very large two-level SOM for the browsing of newsgroups. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'96)*, Bochum, Germany, 1996.
25. K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In *Proc of the Int'l Conf on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, 1996.
26. X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proc of the ACM SIGIR Int'l Conf on Research and Development in Information Retrieval (SIGIR'91)*, Chicago, IL, 1991.
27. D. Merkl. A connectionist view on document classification. In *Proc of the Australasian Database Conf (ADC'95)*, Adelaide, SA, 1995.
28. D. Merkl. Content-based document classification with highly compressed input data. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'95)*, Paris, France, 1995.
29. D. Merkl. Content-based software classification by self-organization. In *Proc of the IEEE Int'l Conf on Neural Networks (ICNN'95)*, Perth, WA, 1995.
30. D. Merkl. The effect of lateral inhibition on learning speed and precision of a self-organizing map. In *Proc of the Australian Conf on Neural Networks*, Sydney, NSW, 1995.
31. D. Merkl. Exploration of document collections with self-organizing maps: A novel approach to similarity representation. In *Proc of the European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'97)*, Trondheim, Norway, 1997.
32. 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.
33. D. Merkl. Text classification with self-organizing maps: Some lessons learned. *Neurocomputing*, 21(1–3), 1998.
34. D. Merkl and A. Rauber. Uncovering associations between documents. In *Proc of the IJCAI'99 Workshop on Text Mining*, Stockholm, Sweden, 1999.

35. D. Merkl, E. Schweighofer, and W. Winiwarter. CONCAT: Connotation analysis of thesauri based on the interpretation of context meaning. In *Proc of the Int'l Conference on Database and Expert Systems Applications (DEXA'94)*, Athens, Greece, 1994.
36. R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2, 1990.
37. R. Miikkulainen. Self-organizing process based on lateral inhibition and synaptic resource redistribution. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'91)*, Espoo, Finland, 1991.
38. R. Miikkulainen. *Subsymbolic Natural Language Processing: An integrated model of scripts, lexicon, and memory.* MIT-Press, Cambridge, MA, 1993.
39. A. Rauber and D. Merkl. Creating an order in distributed digital libraries by integrating independent self-organizing maps. In *Proc of the Int'l Conf on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.
40. A. Rauber and D. Merkl. Automatic labeling of self-organizing maps: Making a treasure–map reveal its secrets. In *Proc of the Pacific Asia Conf on Knowledge Discovery and Data Mining (PAKDD'99)*, Beijing, China, 1999.
41. H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61, 1989.
42. D. E. Rose. *A Symbolic and Connectionist Approach to Legal Information Retrieval.* Lawrence Erlbaum, Hillsdale, 1994.
43. D. E. Rose and R. K. Belew. Legal information retrieval: A hybrid approach. In *Proc of the Int'l Conference on Artificial Intelligence and Law (ICAIL'89)*, Vancouver, Canada, 1989.
44. D. Roussinov and M. Ramsey. Information forage through adaptive visualization. In *Proc of the ACM Int'l Conf on Digital Libraries (DL'98)*, Pittsburgh, PA, 1998.
45. G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA, 1989.
46. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 1988.
47. H. R. Turtle and W. B. Croft. A comparison of text retrieval models. *Computer Journal*, 35(3), 1992.
48. R. Wilkinson and P. Hingston. Incorporating the vector space model in a neural network used for information retrieval. In *Proc of the ACM SIGIR Int'l Conf on Research and Development in Information Retrieval (SIGIR'91)*, Chicago, IL, 1991.
49. P. Willet. Recend trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24, 1988.
50. J. Zavrel. Neural navigation interfaces for information retrieval: Are they more than an appealing idea? *Artificial Intelligence Review*, 10, 1996.