Diplomarbeit

The Growing Hierarchical Self-Organizing Map: Uncovering Hierarchical Structure in Data

ausgeführt am Institut für
Softwaretechnik
der Technischen Universität Wien

unter Anleitung von
ao. Univ. Prof. Dr. Dieter Merkl
und
Univ. Ass. DI Andreas Rauber

durch

Michael Dittenbach
Schenkendorfgasse 14–16/2/12
A–1210 Wien

Wien, am 7. Dezember 2000

Unterschrift

Abstract

Discovering the inherent structure in data has become one of the major challenges in data mining applications. It requires stable and adaptive models that are capable of handling the typically very high-dimensional feature spaces, with current approaches hardly incorporating these requirements within a single model.

In this thesis we present the *Growing Hierarchical Self-Organizing Map* (GHSOM), a neural network model based on the *Self-Organizing Map*. The main feature of this novel architecture is its capability of growing both in terms of map size as well as in a three-dimensional tree-structure in order to represent the hierarchical structure present in a data collection during an unsupervised training process. This capability, combined with the stability of the *Self-Organizing Map* for high-dimensional feature space representation, makes it an ideal tool for data analysis and exploration.

We demonstrate the potential of the GHSOM with an application from the information retrieval domain, which is prototypical both of the high-dimensional feature spaces frequently encountered in today's applications as well as of the hierarchical nature of data.

Contents

List of Figures

List of Tables

Acknowledgments

The work for this thesis was carried out at the Department of Software Technology at the Vienna University of Technology. I'd like to thank all the people of the department for a pleasant time, nice coffee breaks, and interesting discussions.

Very special thanks go to my supervisor Dieter Merkl, who gave me a lot of support during the work on this thesis and yet more. He helped me a lot in writing scientific papers [Dit00a, Dit00b] and encouraged me in presenting them at international conferences.

Furthermore, I am indebted to Andreas Rauber who *endured* my presence in his office, where we spent a lot of hours (often late into the night) working, discussing, and thinking about new ideas. We worked together on [Rau00], the publications mentioned above, and enjoyed a nice week in Slovakia where [Dit00] was presented at a workshop.

Additional thanks go to Samuel Kaski from the Helsinki University of Technology, who gave me helpful hints on scientific writing, too.

Chapter 1

Introduction

"Well, for starters I'll have
who, what, when and where,
and then whither, whether, whence
and where for the follow –
and one big side order of why."
– Douglas Adams' "The Hitchhiker's Guide
to the Galaxy Radio Series"

DATA mining, or more generally, pattern recognition and knowledge acquisition, depend heavily on suitable unsupervised learning methods. The function of these methods is to develop an optimal partitioning, i.e. clustering, of the data set to be analyzed. Cluster analysis is the organization of a collection of patterns (usually represented as vectors of measurements, or points in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster [Jai99]. In other words, the objective of unsupervised learning methods in data mining applications is to identify groupings in an unlabeled set of data vectors that share some semantic similarities. This helps the user to build a cognitive model of the data, thus fostering the detection of the inherent structure and the interrelationship of data. However, in many applications little to no prior information about underlying models for the data is available. In such a situation clustering provides a particularly appropriate approach to the analysis of data.

The *Self-Organizing Map* (SOM) [Koh82] is an artificial neural network model that has shown to be well-suited for mapping high-dimensional data into a two-dimensional representation space. This non-linear mapping preserves distances between input data, such that similar data will be represented spatially close in the output space. This feature proved to be exceptionally successful for clustering [Koh95], visualization and analysis of high-dimensional data. A bibliography of over 3000 research papers about the SOM and a wide variety of its applications can be found in [Kas98a].

Especially the utilization of the SOM for information retrieval purposes in large document collections has gained interest in the last few years [Lin91, Lag96, Mer97a, Rau98]. However, with the increasing amount of information available in today's society some limitations of the SOM have to be addressed. One of these disadvantages is its fixed size in terms of the number and the particular arrangement of the neural processing elements. Without a priori knowledge about the type and the organization of the data it is difficult to predefine the network's size in order to reach satisfying results.

Another drawback is the flat structure of the SOM although many data collections can be considered hierarchically structured. Text document collections, for example, can be decomposed into several major subjects which can be divided again into multiple sub-topics and so on. Hence, we propose a novel neural network architecture, namely the *Growing Hierarchical Self-Organizing Map* (GHSOM). Its adaptive architecture combines the advantages of several variants of the *Self-Organizing Map.* This model consists of several layers of independent growing *Self-Organizing Maps.* During the training process the size of these maps and the structure of the hierarchy are determined dynamically to resemble the structure of the input data as accurately as possible.

To stress the point of hierarchical organization, consider a conventional library building as depicted in Figure 1.1. The books and other documents are usually located at a very granular level in a bookshelf, the bookshelves are positioned somewhere in a room according to the topics they cover. Finally, the rooms are spread onto several floors in the building. Therefore, we believe that a hierarchical structure of *Self-Organizing Maps* on which the documents are organized based on their content is a very convenient way for browsing in a document archive.

Figure 1.1: **Library of Exeter:** The way documents have been organized for centuries in conventional libraries.

However, the hierarchical content-based classification performed by the GHSOM algorithm should not be mistaken for the classification of documents present in conventional libraries, because in libraries, usually, not only the content is determining for the position of a document.

We show the usefulness of the *Growing Hierarchical Self-Organizing Map* with an application in document archive organization. Document archives represent a convenient application scenario because they are, by their very nature, represented as high-dimensional data. In particular, we show the results from two experiments. The first one is based on the *TIME Magazine* collection. This collection comprises 420 articles from the *TIME Magazine*, covering a variety of topics ranging from international politics to social gossip. Two different hierarchical structures will be compared. The second experiment is based on a much larger document collection of more than 10,000 articles from the daily Austrian newspaper *Der Standard.*

The remainder of this thesis is organized as follows. The neural network models, such as the *Self-Organizing Map*, *Growing Grid* and the *Hierarchical Feature Map*, vitally important for the understanding of the GHSOM, are explained in Chapter 2. Then, a detailed description of the training process of the GHSOM follows in Chapter 3. An example from the text classification domain is presented in chapter 4 where the description of a method for representing text documents as high-dimensional vectors is given along with an approach to describe the maps of a trained GHSOM. Two text document collections organized by our architecture are presented with detailed descriptions of the represented topical hierarchies. Finally, we present some conclusions and suggestions for further research in Chapter 5.

Chapter 2

The Self-Organizing Map and Some of its Variants

When we write programs that "learn",
it turns out that we do and they don't.
– Alan J. Perlis' "Epigrams in Programming"

THIS chapter contains the description of several neural network models which are necessary for the understanding of the *Growing Hierarchical Self-Organizing Map*, the GHSOM for short. First, the *Self-Organizing Map* (SOM) [Koh82] is described by means of its structure and its training algorithm. Then, descriptions of two enhanced models follow, which are based on the SOM and show certain characteristics which are advantageous for our neural network architecture. These favorable features will be combined in the design of the GHSOM. One of these models is the *Growing Grid* [Fri95] which is a rectangular-shaped *Self-Organizing Map* that grows during the training process by insertion of rows or columns of units. The second enhanced SOM derivative is the *Hierarchical Feature Map* [Mii90] which consists of a multi-layered hierarchy of independent fix-sized *Self-Organizing Maps.*

2.1 The Self-Organizing Map

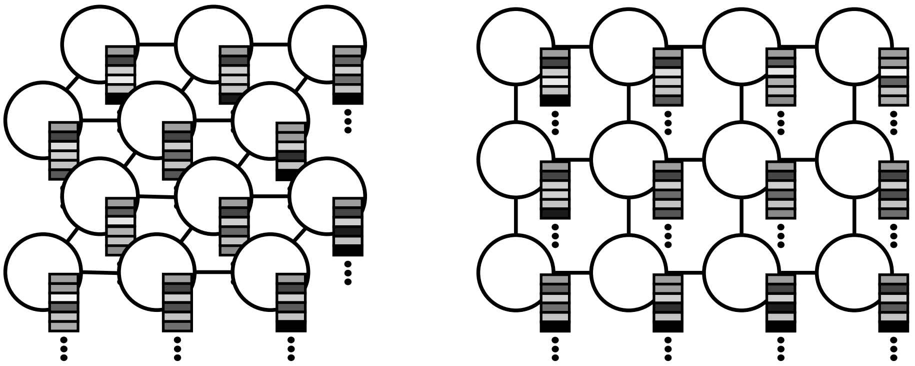
The *Self-Organizing Map* (SOM), as proposed in [Koh82] and described thoroughly in [Koh89, Koh95] is a well known representative of unsupervised artificial neural networks especially in the fields of clustering, data classification and data visualization. It performs a non-linear projection of high-dimensional data onto a usually two-dimensional map preserving the topology of the input space as faithfully as possible, i.e. similar input patterns will be mapped onto spatially close regions in the output space. As a consequence, the relationship between input data is mirrored in terms of the distance of the respective representatives in the output space. Thus, the SOM is a convenient tool for the visualization and the exploration of high-dimensional data.

2.1.1 Architecture

The input data are represented by n-dimensional vectors $x_j \in \Re^n$, i.e. the data $x_j = (x_{j_1}, x_{j_2}, \ldots, x_{j_n})^T$ are described by n features in the input space. The set of input vectors x_j for a SOM will be denoted as I. The *Self-Organizing Map* consists of an input layer which propagates the input data in parallel to a number of neurons (units) in the output layer which may be organized in a hexagonal (Fig. 2.1(a)), rectangular (Fig. 2.1(b)) or even irregular, usually two-dimensional, lattice. A rectangular layout will be assumed for the rest of this description. Every unit i has an associated weight vector $m_i = (m_{i_1}, m_{i_2}, \ldots m_{i_n})^T$ of the same dimensionality n as the input vectors. In Figure 2.1, the weight vectors are represented by arrays of boxes in different shades of gray according to the respective values of the weight vector components. These weight vectors may either be initialized randomly, with random samples from the input data set or by more sophisticated methods such as, for example, *Principle Component Analysis* [Hot33].

2.1.2 Training Algorithm

In the following equations, we make use of a discrete time notation, with t denoting the current training iteration. The training starts by random selection of an input vector x_j. Then, the unit c with the smallest distance between its assigned weight vector m_c and input x_j in the Euclidean space is selected as the *best-matching unit* (hereafter referred to as *winner*) according to Equation 2.1. The Euclidean distance is denoted as $\| \cdot \|$.



(a) Hexagonal Lattice

(b) Rectangular Lattice

Figure 2.1: **Different Types of Neuron Arrangements:** The units of aSOM may be arranged in different types of lattices. An n-dimensional weight vector is assigned to every unit, depicted by an array of shaded boxes which represent the different values of the weight vector components.

$$\begin{aligned} c &= \arg\min_i(\|x_j - m_i\|) = \\ &= \arg\min_i\left(\sqrt{\sum_{k=1}^{n}(x_{j_k} - m_{i_k})^2}\right) \end{aligned} \tag{2.1}$$

In other words, the input x_j is represented best by unit c. To increase the probability for this unit to be chosen as *winner* if the same input is selected in subsequent training iterations, the difference between the unit's weight vector m_c and input x_j will be decreased. This gradual adaptation of the weight vector is controlled by the *learning rate* parameter $\alpha(t) \in [0, 1]$. It usually is a time-decreasing function with $\lim_{t\to\infty}\alpha(t) = 0$. Hence, weight vectors will be adapted stronger at the beginning of the training process. A rather low value of $\alpha(t)$ at the end of the training process leads to a fine-tuning phase.

To achieve a topology preserving mapping, i.e. preserving similarity relations between input data on the output space, not only the weight vector of the *winner* c will be adapted, but also the weight vectors of units in its vicinity. Thereby, input data similar to x_j are more likely to be represented in the region of the

SOM where the *winner* is located. The adaptation strength $h_{ci}(t)$ of neighboring units is determined by their distance from unit c on the *Self-Organizing Map*. This is a time-decreasing function as given in Equation 2.2 (Gaussian function).

$$h_{ci}(t) = \exp - \frac{\|r_c - r_i\|^2}{2\sigma^2(t)} \tag{2.2}$$

where $r_c \in \Re^2$ is the location vector of the winning unit c on the grid and $r_i \in \Re^2$ the location of a neighboring unit i. Parameter $\sigma(t)$ is the time-dependent factor. It can be seen from Equation 2.2 that units closer to the winner are adapted stronger than units which are farther away. A high value of h_{ci} at the beginning of the training process leads to a global organization of the units' weight vectors, i.e. neighboring units have similar weight vectors. By decreasing the neighborhood function successionally in the course of time, the adaptations become more local.

A simpler neighborhood function is, to define a set of units $N_c(t)$ (*neighborhood kernel*) around *winner* c at time t, whereby the adaptation strength h_{ci} of the neighboring units is determined as follows:

$$h_{ci}(t) = \begin{cases} \alpha(t) & \text{if } i \in N_c(t), \\ 0 & \text{if } i \notin N_c(t). \end{cases} \tag{2.3}$$

Hence, only weight vectors of units within the *neighborhood kernel* are adapted. The computational load during training a large map can benefit from this approach, because only a subset of the units require weight vector adaptation. Whereas, with the previously described Gaussian function (see Eq. 2.2), every unit's weight vector has to be adapted at every training iteration. Again, a rather large neighborhood kernel, say, half the diameter of the network, at the beginning of the training process is required to reach a globally ordered map.

Having defined the learning rate $\alpha(t)$ and the neighborhood function $h_{ci}(t)$, the weight vector $m_i(t+1)$ of a unit i is adapted by adding a portion $\alpha(t) \cdot h_{ci}(t)$ of the vector difference $[x(t) - m_i(t)]$ to $m_i(t)$ according to Equation 2.4. $x(t)$ denotes the current input vector at time t of the set of input vectors, $x \in I$.

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \tag{2.4}$$

As a consequence of Eq. 2.4, the weight vector of the winner and the weight vectors of the units in its vicinity are "moved" towards the input vector. Hence, its more likely that this and similar input vectors are mapped into this very region of the map in successional learning iterations.

Figure 2.2 illustrates a SOM along with a graphical representation of the Gaussian neighborhood function. On the left-hand side, the input space $\Re^n$ is depicted. We find the location of the *winner's* weight vector $m_c(t)$ at time t and of the current input vector x in this input space. The weight vector of the *winner* $m_c(t+1)$ after the adaptation at time $t+1$ can be found closer to the input vector. The movement is represented by the solid arrow.

On the right-hand side of Figure 2.2, the *Self-Organizing Map* is depicted by an array of differently shaded circles. A dotted arrow shows the relation between the winning unit c and its weight vector in the input space. The different shades of the units represent the different adaptation strengths according to the distance from the *winner*. The darker the unit, i.e. the closer it is to the *winner*, the stronger its weight vector is adapted.

In short, one iteration of the SOM training algorithm can be summarized as follows:

1. random selection of an input vector x

2. search for best-matching unit (Eq. 2.1)

3. weight vector adaptation of the winner and its neighbors (Eq. 2.4)

4. modification of learning rate and neighborhood range

Now, that the first iteration is finished, the training process proceeds with the selection of the next input vector and continues until a predefined number of training steps or another stopping criterion is met. For example, a certain mathematical quality criterion could serve as a condition for terminating the training process. Otherwise, the training could be terminated, if a stable organization of the input vectors within the two-dimensional lattice is reached.

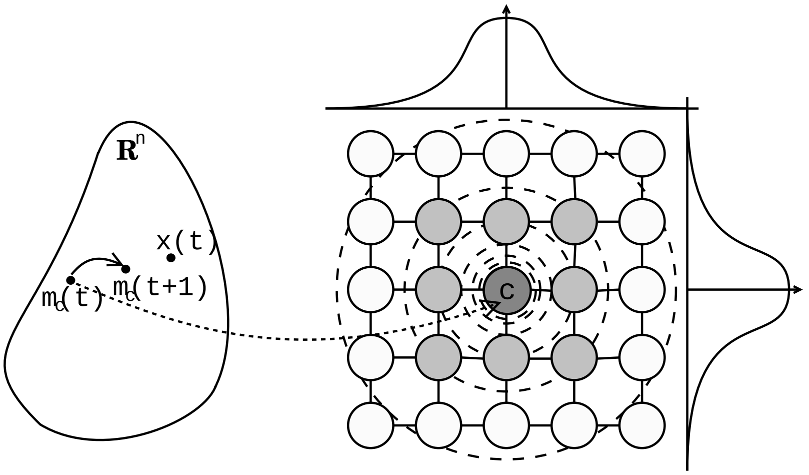


Figure 2.2: SOM **Lattice:** The most common topological neighborhood is a (2-dimensional) Gaussian function. The adaptation strength of the individual units is indicated by different shades of gray and by the horizontal and vertical Gaussian.

2.1.3 Discussion

A large number of publications have been presented in which the *Self-Organizing Map* has been used as a tool for a wide variety of applications [Kas98a]. Especially in the field of data classification, clustering and information retrieval [Lin91, Lag96, Hon97, Köh96, Rau98], just to mention a few, a lot of work has been done. The mapping of high-dimensional data onto a 2-dimensional grid preserving the topology of the input space provides a convenient interface for exploratory data analysis [Kas96].

One of the disadvantages of this artificial neural network is that the size of the map has to be determined in advance. One can estimate the map size by calculating how many data vectors have to be mapped on a node on average given the number of input vectors, but without further knowledge about the distribution of the data this size may be inadequate because of one or more accumulation points of the data. Such a case would lead to a map with empty regions and other parts where a massive amount of data would be mapped although the SOM exhibits a kind of magnification characteristic, such that more prominently present data vectors are automatically distributed over a larger region of the map.

Another shortcoming of the SOM is that large maps are more sensitive to the parameter settings [Koh95], i.e. the initial values of $\sigma(t)$ and $\alpha(t)$ may be crucial for the global organization of the data and the convergence of the algorithm. Also the computational load increases with the size of the map, because of the possibly large neighborhood radius more weight vector adaptations have to be performed.

Furthermore, notwithstanding the good mapping quality achieved with the SOM algorithm, the cluster boundaries are not implicitly visible. Clusters may not even be disjoint, i.e. some data, actually belonging to a certain class A, may overlap with a different cluster B of data. Hence, a method has to be applied after the training process to facilitate the visual exploration of the map. Some of the approaches, mainly coloring techniques, will be outlined in Section 4.2.

2.2 Growing Grid

To overcome the drawback of defining the network size prior to the training process, several adaptive models have been proposed [Bau97, Bla93, Fri94, Fri95, Fri96]. We are especially interested in the *Growing Grid* (GG) model introduced by Fritzke in [Fri95]. It is a self-organizing neural network model based on the SOM with a rectangular grid structure which is generated by a growth process to adapt the size of the network according to the properties of the input space. Initially, the size of this network is rather small. During the training process, ever after a fixed number of training iterations, rows or columns of units are inserted. The decision, at which location of the grid rows or columns are to be inserted, is made by examining the number of times the various units have been chosen as *winner*. A high value of a unit's *winner counter* indicates that a large number of input vectors have been mapped onto this very unit. Therefore, the representation space will be expanded by inserting more units in this units immediate vicinity to provide more map space for the distribution of the input vectors. The weight vectors of these new units are initialized according to their neighbor's weight vectors to preserve the already achieved ordering of the input vectors.

Another difference between the training processes of the *Self-Organizing Map* and *Growing Grid* is the constant neighborhood function and adaptation strength. During SOM training, both of these variables are decreasing in the course of time. Due to the growth of the GG network the constant neighborhood is automatically decreasing relative to the size of the network. The constant learning rate during the growth phase prevents the algorithm from converging into a stable state too early. A fine-tuning phase with decreasing adaptation strength leads to a stable organization of the input vectors on the map. An application-specific stopping criterion or the maximum number of units may be defined to terminate the training process.

2.2.1 Architecture and Training Algorithm

Initially, the rectangular-shaped grid M consists of a rather small number $p \times q$ of units where the width p and height q of the map must have at least the value of two:

$$M = [m_{ij}], \qquad 2 \leq i \leq p, \qquad 2 \leq j \leq q \tag{2.5}$$

Please note, for the sake of readability we further refer to the units' weight vectors by using a single index, e.g. m_i, denoting the position in the grid. As described above (see Section 2.1), a weight vector m_i, with the same dimensionality n as the input vectors, is assigned to each unit i. Additionally, every unit has an associated resource variable η_i which we will further refer to as *winner counter* of unit i, because it holds the number of times a unit has been selected as the best-matching unit. This information is used for the decision where to insert a new row or column of units. Again, the set of input vectors is denoted as I and a single input vector presented to the map as x_j.

The winner selection and the weight vector update is similar to the standard SOM training algorithm except that the *city-block metric* (cf. Eq. 2.6) is used for calculating the distance between two units on the 2-dimensional map. The distance $d(p, q)$ between unit p and unit q is calculated as the sum of the differences of their respective location vector components.

$$d(p,q) = |p_1 - q_1| + |p_2 - q_2| \tag{2.6}$$

The neighborhood function, which makes use of this metric, is a Gaussian function with constant width. Also the learning rate remains unchanged during the growth phase of the training process. However, every time a unit is chosen as the best-matching unit c due to the smallest Euclidean distance between its weight vector and the input vector $||m_c - x_j||$, the *winner counter* η_c is incremented by 1, as given in Equation 2.7.

$$\eta_c = \eta_c + 1 \tag{2.7}$$

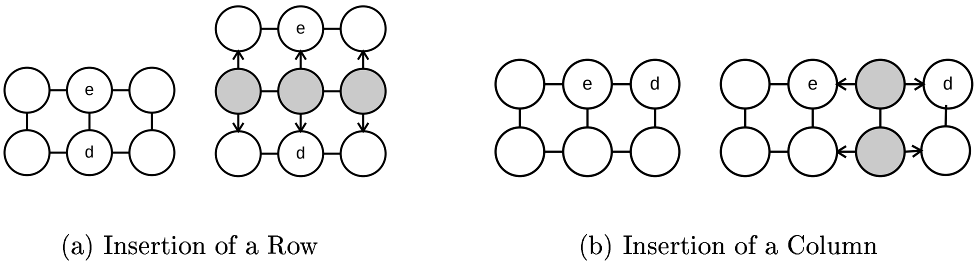
After a fixed number $p \times q \times \lambda$ of training iterations new units are inserted. The parameter λ signifies the average number of adaptation steps per unit on a grid of size $p \times q$. A row or column of units is inserted to that position of the grid where many input vectors are being mapped onto in order to distribute these inputs more evenly. The unit e with the highest *winner counter* (see Expression 2.8) indicates a possibly high number of represented input vectors.

$$e = \arg\max_i(\eta_i) \tag{2.8}$$

A better mapping of the input data in this area can be achieved by increasing the number of units in between e and one of its directly neighboring units. The set of directly neighboring units is further referred to as N_e. The direct neighbor d which has assigned the most dissimilar weight vector m_d with respect to m_e is chosen, because a high variance of the input data implicates the need for an enhanced resolution at this very location. Unit d of the set of neighbors N_e is determined as follows:

$$d = \arg\max_{n \in N_e}(||m_e - m_n||) \tag{2.9}$$

Consider Figure 2.3 as a graphical representation of a *Growing Grid*. On the left-hand side of Figure 2.3(a), the map is shown before and on the right-hand



(a) Insertion of a Row

(b) Insertion of a Column

Figure 2.3: **Insertion of Units:** A row (a) or a column (b) is inserted in between unit e with the highest *winner counter* and the neighboring unit d with the largest distance between its weight vector and the weight vector of e in the Euclidean space.

side after the insertion of a row of units. New units are depicted as shaded circles. Unit e and its most dissimilar neighbor d are located in the same column, thus a row is inserted in between row 1 and 2. Figure 2.3(b) depicts the analogous case, where unit e and unit d are situated in the same row.

The weight vector elements m_{g_i} for a new unit g are interpolated as the mean of the respective elements m_{n1_i} and m_{n2_i} of the two neighbors' weight vectors m_{n1} and m_{n2} (Expression 2.10). The arrows in Figure 2.3 point to the respective neighboring units. With this initialization, the input vectors are spread out more evenly over the units in this particular area and the information already gained is not distorted as it would be, by using randomly initialized weight vectors.

$$m_{g_i} = \frac{1}{2} \cdot (m_{n1_i} + m_{n2_i}) \tag{2.10}$$

Then all *winner counters* are reset (Equation 2.11).

$$\eta_i = 0, \qquad \forall i \in M \tag{2.11}$$

and the adaptation process continues until a stopping criterion described below is met.

There exist several criteria which can be used to terminate the training process. The simplest stopping criterion might be to define a maximum number of nodes on the map, but more sophisticated and application specific conditions for terminating the growth process may be advantageous.

The final part of the training process of the *Growing Grid* is a fine-tuning phase in which the quality of the mapping is further increased and the locations of the mapped input vectors are settled. This is achieved by training the map for a certain number of iterations with time-decreasing learning rate to converge into a stable state, because this is inhibited by the constant learning rate during the growth phase. During this fine-tuning phase, however, no further units are added to the network.

2.2.2 Discussion

The *Growing Grid* architecture overcomes the disadvantage of the SOM of having a fixed network size. However, the lattice is limited to a rectangular shape. This model adapts its size during the training process according to the structure of the input space which is an important feature, because a priori information about the data may be unavailable in many cases. Other advantages are [Fri96]:

1. the possibility of using problem specific criteria to determine the location where units are inserted (e.g. a topographic function [Bau97]),

2. eventual interruption and continuation of the training process at a later time because there are no different phases during self-organization due to constant parameters and

3. fewer parameters to define. Nevertheless, a large number of input vectors can only be displayed with a high granularity, if many rows and columns are inserted during training. Hence, we are still faced with the drawback of having large maps which are complex and difficult to survey.

2.3 Hierarchical Feature Map

The *Hierarchical Feature Map* (HFM) has been introduced in [Mii90] for the use in a script recognition system. The idea is to show the hierarchical taxonomy of scripts, tracks and role bindings extracted from script-based stories by using

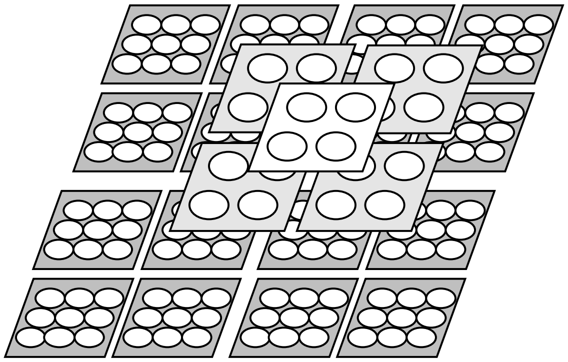


Figure 2.4: **Hierarchical Feature Map:** A balanced tree of several layers of *Self-Organizing Maps*.

a hierarchical system of fix-sized *Self-Organizing Maps*, such that different levels of representations can be made visible. Furthermore, the training task is divided into hierarchical subgoals and the computation time can be reduced effectively. A comparison of the quality of data representation and of computational times needed for the training of a specific input data set with both the SOM and the HFM can be found in [Mer97a, Mer97b].

2.3.1 Architecture and Training Algorithm

As noted above the HFM architecture consists of several layers of independent *Self-Organizing Maps*. For each unit in a layer of the hierarchy a new map is added to the next layer (Fig. 2.4). The architecture can be seen as a balanced tree of fix-sized SOMs with a certain predefined depth. The training process for every single *Self-Organizing Map* is equal to the SOM algorithm described in Section Sect. 2.1.

The training of the HFM is performed sequentially from the first layer downwards along the hierarchy. As soon as the training process of the first level map is finished every map in the second layer is trained, but only with the subset of the data which has been mapped onto the corresponding unit in the first-layer map. In addition, the dimensionality of the input vectors for a second-layer map can be reduced by omitting the similar features for the respective set of vectors. This dimensionality reduction is based on the consideration that a subset of the data, which has been mapped onto one unit of a map, has a certain number of features in common. These features are not useful for describing the differences of this subset of the data at a subsequent layer of the hierarchy. Hence, only dissimilar features are decisive for the organization of the data at a more granular view.

The training procedure is applied to all maps in subsequent layers of the hierarchy until the maps in the last layer are trained.

The map in the first layer of the *Hierarchical Feature Map* exhibits a very general organization of the input data due to its usually small size. The respective subsets of the data are trained on independent maps in the lower layers and organized with a higher granularity of data representation. Thus, the data are separated into a hierarchical system of clusters in different gradations.

2.3.2 Discussion

The HFM is an adequate neural network model for representing data which is inherently structured hierarchically. The data exploration is facilitated by dividing the data into multiple subsets represented by different *Self-Organizing Maps* which are rather small and convenient to survey. Also the cluster boundaries are inherently present due to the very structure of this neural network architecture. Moreover, this model leads to a considerable reduction of time needed for training as compared to the SOM. The computational speed-up can be explained by several characteristics of the HFM:

1. The first-layer map is usually small and therefore fast to train.
2. Maps in further layers may be larger, but they have to be trained only with subsets of the data.
3. The dimensionality of these subsets can be reduced by eliminating vector elements which are almost equal.

However, the size of the architecture, i.e. the size of the maps and the depth of the hierarchy, has to be defined in advance, which remains a difficult task if no knowledge about the structure of the data is available. Additionally, the number of categories, sub-categories and further divisions is determined by the architecture instead of having the architecture adapt itself according to the structure of the input data. This problem is discussed in the next chapter where a novel neural network model is introduced to eliminate this drawback.

Chapter 3

The Growing Hierarchical Self-Organizing Map

In a hierarchical organization,
the higher the level,
the greater the confusion.
– Cpt. Ed Murphy "Dow's Law from Murphy's Laws"

IN this chapter we describe the architecture of the *Growing Hierarchical Self-Organizing Map* [Dit00b, Dit00a, Mer00], GHSOM for short, as well as the training algorithm of this adaptive neural network model. A detailed description of the initial structure, the growth and learning process of the hierarchy and the resulting organization of the hierarchically interconnected, growing SOMs is presented.

3.1 The Key Idea

The basic idea behind the *Growing Hierarchical Self-Organizing Map* (GHSOM) is to represent the inherent hierarchical structure present in many data collections in the most accurate way. As a result of the training process, a representation reflecting this hierarchical structure should be provided. Especially text documents, e.g. books, newspaper articles or scientific papers, show the characteristic of being organized in topic hierarchies, i.e. they can be decomposed into a variety of subjects.

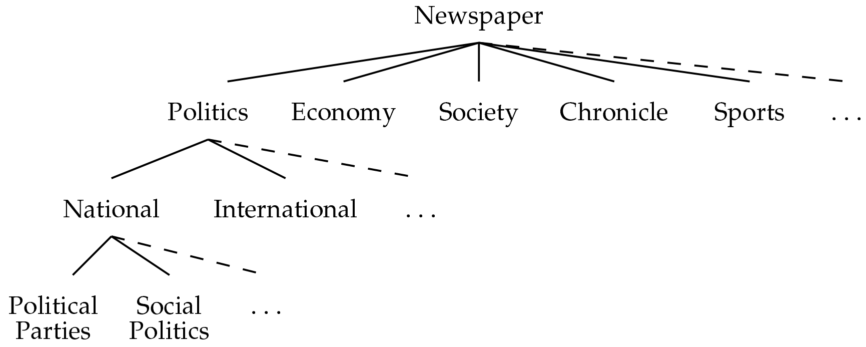


Figure 3.1: **Possible Newspaper Taxonomy:** Different levels of granularity of topics covered by a newspaper.

In Figure 3.1, the main topics of a newspaper can be identifies as, for example, politics, economy, society, chronicle, and sports. Political issues may further be divided into national and international political news. Hence, a document collection viewed in different levels of granularity will exhibit different numbers of topics respectively.

Similar to the *Hierarchical Feature Map* we use independent *Self-Organizing Maps* on each level of the hierarchy. However, these SOMs grow in size analogously to the *Growing Grid*, but with a different unit insertion strategy. The GHSOM architecture consists of only one growing SOM at the beginning of the training process. Contrary to the HFM, maps are added to subsequent layers during training to represent input vectors in more detail only if needed. Hence, the resulting hierarchy is a tree-like structure of SOMs which is not necessarily balanced in terms of equal depth of the branches.

A priori knowledge about the organization of data collections is not given in many cases. Some topics may be more dominant than others and appropriately predefining the shape of a static hierarchy, i.e. the depth of different branches or the sizes of the *Self-Organizing Maps*, would be a very difficult task. To overcome the limitations of the neural network models presented in Chapter 2 we developed the GHSOM, which dynamically fits its multi-layered architecture according to the structure of the data.

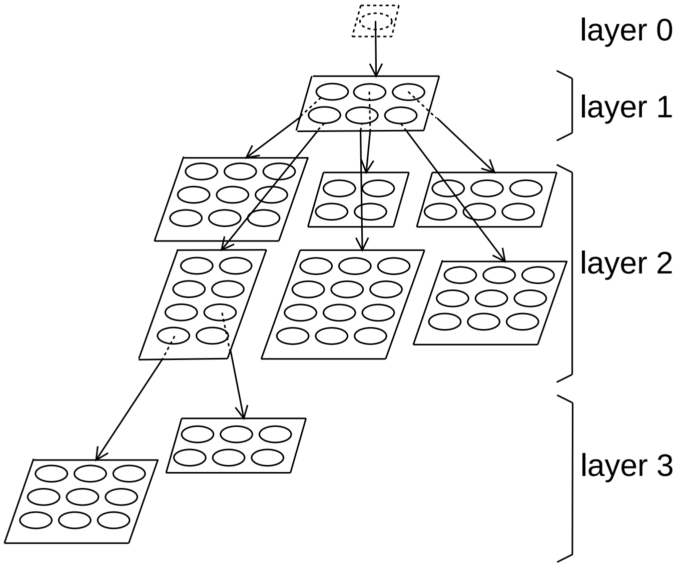


Figure 3.2: **Trained GHSOM:** The units of a 3 × 2 SOM in the first layer of the hierarchy have been expanded into the second layer. The data mapped onto two units of a layer 2 map are explained in more detail on two layer 3 maps.

3.2 Architecture

The GHSOM has a hierarchical structure of multiple layers where each layer consists of several independent growing *Self-Organizing Maps*. A graphical representation of a trained GHSOM is given in Figure 3.2. The map in layer 1 consists of 3 × 2 units and provides only a rough organization of the main clusters in the data. The six independent maps in the second layer offer a more detailed view of the data. The input data for one map is the subset which has been mapped onto the corresponding unit in the upper layer. Two units of a second-layer map have further been expanded into the third layer of the hierarchy.

It has to be noted that the maps have different sizes according to the structure of the data. The dynamic growth of the GHSOM relieves us from the burden of predefining the structure of the architecture. The layer 0 is necessary for the control of the growth process and will be explained later in Sections 3.2.3 and 3.3.

3.2.1 Growing Self-Organizing Maps

The maps used in the GHSOM are a growing variant of the SOM. Due to the possibly diverse characteristics of different parts of the data we believe that growing SOMs are well-suited to reflect these diversities. The representation spaces, i.e. map sizes, for the different clusters discovered in the data should be dynamically adapted during training.

The main difference between the growing *Self-Organizing Maps* used in the GHSOM and the *Growing Grid*, described in Section 2.2, is that unit insertion is guided by means of the achieved quality of data representation instead of the *winner counter*. After a fixed number $\kappa \times \lambda$ training iterations (i.e. one training cycle) the unit with the largest average deviation between its weight vector and the input vectors mapped onto this very unit is selected as the *error unit*, where κ is the number of input vectors for this map. Consequently, λ denotes the average number of presentations of an input vector during one training cycle.

The criterion for the selection of the *error unit* is called *mean quantization error* (cf. Equation 3.1) of a unit, further referred to as *mqe*. The *mean quantization error* mqe_i of a unit i is calculated as the mean Euclidean distance between the weight vector m_i and the n input vectors x_j which are elements of the set of input vectors S_i that are mapped onto this unit i:

$$mqe_i = \frac{1}{n} \cdot \sum_{x_j \in S_i} \|m_i - x_j\|, \qquad n = |S_i| \tag{3.1}$$

where $|\cdot|$ denotes the cardinality of a set.

A high *mqe* of a unit indicates either that a large number of input vectors is mapped onto this unit, or that the input vectors are very diverse. In both cases, the data representation quality can be increased by adding units to this particular area of the map. As a consequence of the unit insertion, the input vectors will be spread more evenly over the additional map space.

A new row or column of units is inserted in between the *error unit* and its most dissimilar neighbor. The weight vectors of the new units are initialized as the average of their corresponding neighbors.

More formally, let I be the set of input vectors, $S_i \subseteq I$ a subset of this data which is mapped onto unit i, m_i the weight vector of unit i and x_j an input vector of S_i. Then, the *error unit* e is determined as the unit with the maximum *mean quantization error*:

$$e = \arg\max_i \left(\frac{1}{n} \cdot \sum_{x_j \in S_i} \|m_i - x_j\| \right), \qquad n = |S_i| \tag{3.2}$$

The selection of the most dissimilar neighbor d is performed according to Equation 2.9 where the maximum distance between the weight vector of unit e and the weight vectors of the neighboring units is calculated. A complete row or column of units is inserted in between d and e. Figure 2.3 can be considered again as a graphical representation of the insertion process of our realization of a growing *Self-Organizing Map*.

The stopping criterion, which prevents the growing self-organizing maps used in the GHSOM from growing indefinitely, is based on the *mean quantization error* of the complete map which will be explained below in Section 3.3.

3.2.2 Learning Rate and Neighborhood Range

In our realization of a growing *Self-Organizing Map*, we use a time-varying learning rate as depicted in Figure 3.3. Every time a row or column of units is inserted the learning rate is set back to its initial value LR_{ini}. The value of LR_{ini} is equal for all maps in the current implementation.

Furthermore, the neighborhood kernel is a Gaussian function (see Figure 3.4(b) which is combined with a neighborhood range function depicted in Figure 3.4(a). As described in Section 2.1.2, the adaptation strength for the neighboring units decreases with increasing distance from the winner. Similar to the learning rate, the neighborhood range is decreased $NR_{ini} - 1$ times during one training cycle and is set back to an initial value NR_{ini} every time units are inserted. The neighborhood range delimits the number of units being adapted according to the neighborhood kernel function around the winner. The parameter σ of the Gaussian neighborhood kernel (cf. Equation 2.2) is decreased $NR_{ini} - 1$ times, too.

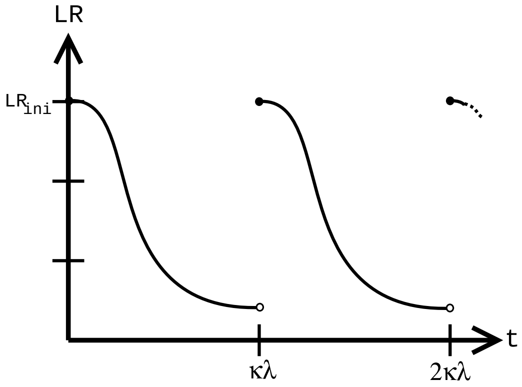


Figure 3.3: **Time-variant Learning Rate:** After every λ training iterations, the learning rate is reset to its initial value LR_{ini}.

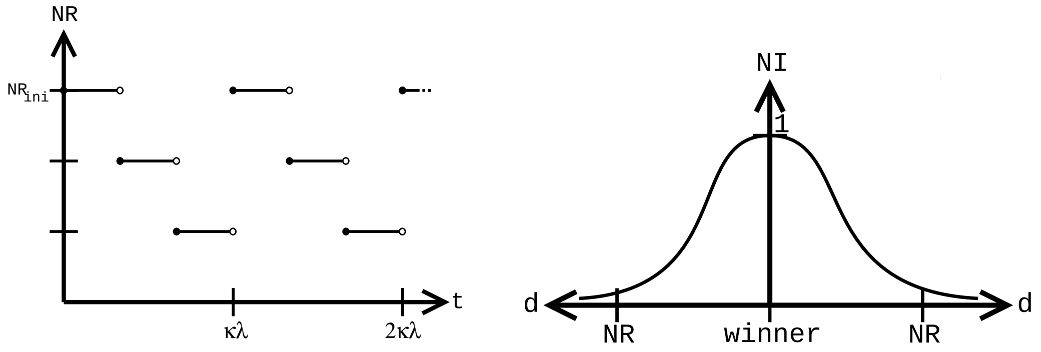
Because of the growth process of the maps used for the GHSOM, the initial neighborhood range narrows relative to the size of the map. Hence, the value of NR_{ini} may be constant during the whole training process. In other words, the constant value of NR_{ini} is decreasing relatively with respect to the increasing size of the map.

Contrary to the *Growing Grid* model where a constant learning rate and neighborhood function was proposed, the fine-tuning phase is integrated in every training cycle between the expansion steps in our variant of a growing *Self-Organizing Map*.

Several experiments with different learning rate functions and neighborhood kernels functions have shown similar results. Thus, for the sake of reduced computational load these functions may be simplified, e.g. linear instead of a Gaussian functions, without noticeable difference in the quality of the results.

3.2.3 Initial Setup

Prior to the training process a “map” in layer 0 consisting of only one unit is created. This unit’s weight vector m_0 is initialized as the average of all input vectors, and its *mean quantization error* is computed as given in Equation 3.3, where n is the number of input vectors x of the set I. The *mean quantization error* of a unit will be referred to as *mqe* in lower case letters.



(a) Time-variant Neighborhood Range: $NR_{ini} = 3$; After every insertion of units, the neighborhood range is set back to its initial value.

(b) Neighborhood Kernel: The adaptation strength of the neighboring units decreases with increasing distance from the winner.

Figure 3.4: **GHSOM Neighborhood:** A combination of time-varying neighborhood range and a Gaussian function defining the adaptation strength of the neighboring units is implemented in the GHSOM training algorithm.

$$mqe_0 = \frac{1}{n} \cdot \sum_{i \in I} \|m_0 - x_i\|, \qquad n = |I| \tag{3.3}$$

The value of mqe_0 can be regarded as a measurement of the dissimilarity of all input data. This measure will play a critical role during the growth process of the neural network, as will be described later. Then a single map in layer 1 is created with a size of, say, 2×2 units.

3.3 Training Algorithm

The first-layer map is trained according to the procedure described in Section 3.2.1 until its *mean quantization error*, referred to as *MQE* in capital letters, reaches a certain fraction τ_1 of the *mqe* of the corresponding unit in the upper layer. The

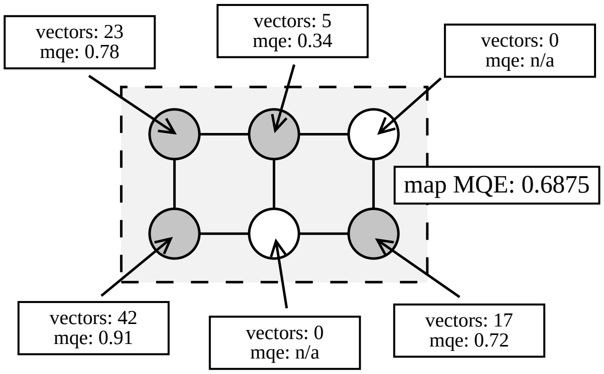


Figure 3.5: **MQE of a Map:** The *MQE* of a growing *Self-Organizing Map* is calculated as the average of the *mean quantization errors* of the units onto which input vectors have been mapped (depicted by shaded circles).

MQE of the map is computed as the mean of all *mean quantization errors* mqe_i (see Eq. 3.1) of the subset U of this maps' units i onto which data are mapped:

$$MQE_m = \frac{1}{u} \cdot \sum_{i \in U} mqe_i, \qquad u = |U| \tag{3.4}$$

To illustrate the point of *MQE* calculation, consider Figure 3.5 as an example of a trained first-layer map. Two of the six units have no input vectors assigned, thus, only the remaining four units are relevant for the calculation the map's MQE.

In case of the first-layer map the stopping criterion for the training process is $MQE_1 < mqe_0 \cdot \tau_1$. Obviously, the smaller the parameter τ_1 is chosen the longer the training will last and consequently the larger the resulting first-layer map will be. In general terms, the stopping criterion for the training of a single map m is defined as:

$$MQE_m < \tau_1 \cdot mqe_u \tag{3.5}$$

where mqe_u is the *mean quantization error* of the corresponding unit u in the upper layer.

If the training of the map is finished, every unit has to be checked for expansion on the next layer. This means, that for units representing a set of too diverse input vectors a new map in the next layer will be created. The threshold for this expansion decision is determined by a second parameter τ_2 which defines the data representation granularity requirement which has to be met by every unit. If Expression 3.6 holds true for unit i, i.e. mqe_i is greater or equal than $\tau_2 \cdot mqe_0$, then a new map in the next layer will be created. The input vectors to train the map with are the ones mapped onto the unit which has just been expanded. Otherwise this unit requires no further expansion.

$$mqe_i \geq \tau_2 \cdot mqe_0 \tag{3.6}$$

Please note, unlike the first criterion determined by τ_1 the second criterion depends on mqe_0, i.e. the *mean quantization error* of the single unit in layer 0, for every unit on all maps in our current implementation. However, further considerations about other (possibly map-dependent) criteria might be made.

Next, all maps in the second layer are trained with the respective subsets of the data. The growth process of a map and the expansion decision for the units, i.e. whether or not additional layers will be constructed, is the same as used for the first-layer map. The whole process is repeated for the subsequent layers as long as no further units require expansion. In other words, the training process terminates as soon as Expression 3.7 holds true for each unit of the lowest layer maps.

$$mqe_i < \tau_2 \cdot mqe_0 \tag{3.7}$$

It should be noted that the training process may not necessarily lead to a balanced hierarchy in terms of all branches having equal depth. As stated in section 3.2, this is one of the main advantages of the GHSOM, because the structure of the hierarchy adapts itself according to the requirements of the input space. Therefore, areas in the input space that require more units for appropriate data representation will create deeper branches than others.

The growth process of the GHSOM is mainly guided by the two parameters τ_1 and τ_2, which merit further considerations.

- τ_2: Parameter τ_2 controls the minimum granularity of data representation, i.e. no unit may represent data at a coarser granularity. If the data mapped onto one single unit still has a larger variation a new map will be added originating from this unit, representing this unit's data in more detail at a subsequent layer.

  This absolute granularity of data representation is specified as a fraction of the inherent dissimilarity of the data collection as such, which is expressed in the *mean quantization error* of the single unit in layer 0 representing all data points. In principle, we could also have chosen an absolute value as a minimal quality criterion. However, we feel that since with most datasets it is difficult to estimate information on its distribution and value ranges without thorough analysis, a percentual, data-driven threshold is more convenient.

  Furthermore, if we decide after the termination of the training process, that a yet more detailed representation would be desirable, it is possible to resume the training process from the respective lower level maps, continuing to both grow them horizontally as well as to add new lower level maps until a stricter quality criterion is satisfied. This parameter thus represents a global termination and quality criterion for the GHSOM.

- τ_1: This parameter controls the actual growth process of the GHSOM. Basically, hierarchical data can be represented in different ways, favoring either (a) shallow hierarchies with rather detailed refinements presented at each subsequent layer, or (b) deep hierarchies, which provide a stricter separation of the various sub-clusters by assigning separate maps.

  In the first case we will prefer larger maps in each layer, which explain larger portions of the data in their flat representation, allowing less hierarchical structuring. As an extreme example we might consider producing a single SOM explaining the complete structure of the data in one single flat map, ignoring all hierarchical information and rather trying to preserve it in the mapping of various clusters on the flat structure.

  In the second case, however, we will prefer rather small maps, each of which

> describes only a small portion of the characteristics of the data, and rather emphasize the detection and representation of hierarchical structure. Basically, the total number of units at the lowest level maps may be expected to be similar in both cases as this is the number of neural processing units necessary for representing the data at the required level of granularity.
>
> Thus, the smaller the parameter τ_1, the larger will be the degree to which the data has to be explained at one single map. This results in larger maps as the map's *mean quantization error* (*MQE*) will be lower the more units are available for representing the data. If τ_1 is set to a rather high value, the *MQE* does not need to fall too far below the *mqe* of the upper layer's unit it is based upon. Thus, a smaller map will satisfy the stopping criterion for the horizontal growth process, requiring the more detailed representation of the data to be performed in subsequent layers.

In a nutshell we can say, that, the smaller the parameter value τ_1, the flatter the hierarchy, and that the lower the setting of parameter τ_2, the larger the number of units in the resulting GHSOM network will be.

Apart from the advantage of automatically determining the number of units required for data representation and the reflection of the hierarchical structure in the data, a considerable speed-up of the GHSOM training process as compared to standard SOM training has to be noted. The reasons for this are two-fold. At the transition from one layer to the next, vector components that are (almost) identical for all input data mapped onto a particular unit can be omitted for the training of the next layer. Shorter input vectors lead directly to reduced training times because of faster winner selection and weight vector adaptation. Secondly, a considerable speed-up results from smaller map sizes, as the number of units that have to be evaluated for winner selection is smaller at each map. This results directly from the fact, that 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.

3.4 Discussion

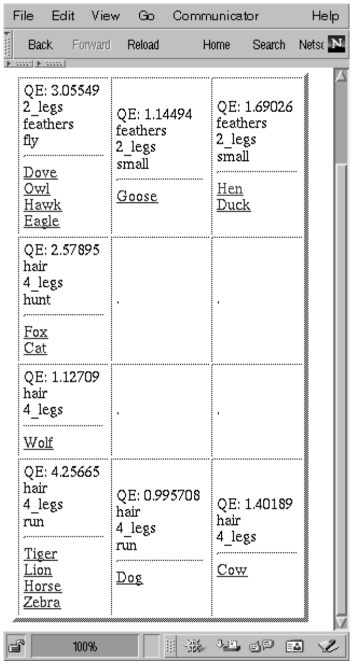
The *Growing Hierarchical Self-Organizing Map* presented in this thesis exploits the advantageous features of the neural network models described in the previous chapter. It is a highly adaptive architecture regarding both the map sizes and the depth of the hierarchy. Especially the adaptive growth process and the possible unbalancedness of a trained GHSOM is reasonable with respect to unevenly distributed input data. The training and growth process is guided solely by the desired data representation granularity. As described in Section 2.3.2, the computational load is also reduced due to the hierarchical structure.

To illustrate the concept of the GHSOM, an example with a toy dataset is given. The dataset consists of 16 animals described by 13 binary features, such as *can_swim*, *can_fly*, *has_hair*, and *has_feathers*. If a feature is true for an animal, the corresponding vector element of the respective input vector is set to 1, otherwise to 0. Two experiments have been carried out. A standard *Self-Organizing Map* is compared to a GHSOM representation of the animal data.

Consider Figure 3.6 as a graphical representation of a 3×4 SOM. In the upper left corner of the map, we find the hunting birds closer to the hunting mammals than the nonhunting birds located to the right. The *dove* is an outlier which has been mapped onto an inappropriate unit.

In the left-most row, the mammals are organized from small (*fox* and *cat*) to the larger ones (*tiger*, *lion*, *horse*, and *zebra*). The *cow* has been mapped onto the unit in the bottom-right corner of the map. Three main clusters present in the data can be identified: nonhunting birds, hunting birds and mammals. The labels for the units describing the common features of the mapped input vectors were obtained using an algorithm detailed in Section 4.2.

The *Growing Hierarchical Self-Organizing Map* representation is depicted in Figure 3.7, with Figure 3.7(a) being the first-layer map. On this map, we find the nonhunting birds in the top-right corner. The two units in the left-hand column have further been expanded to the second-layer. The labels for the unit in the top-left corner of the map hint at small, hunting birds, whereas the labes for the unit in the bottom-left corner indicate mammals being present at the map in the next layer. Figure 3.7(b) shows a more detaile view on the birds. Again,


File Edit View Go Communicator Help
Back Forward Reload Home Search Netsc
100%

QE: 3.05549 2_legs feathers fly Dove Owl Hawk Eagle	QE: 1.14494 feathers 2_legs small Goose	QE: 1.69026 feathers 2_legs small Hen Duck
QE: 2.57895 hair 4_legs hunt Fox Cat	.	.
QE: 1.12709 hair 4_legs Wolf	.	.
QE: 4.25665 hair 4_legs run Tiger Lion Horse Zebra	QE: 0.995708 hair 4_legs run Dog	QE: 1.40189 hair 4_legs Cow

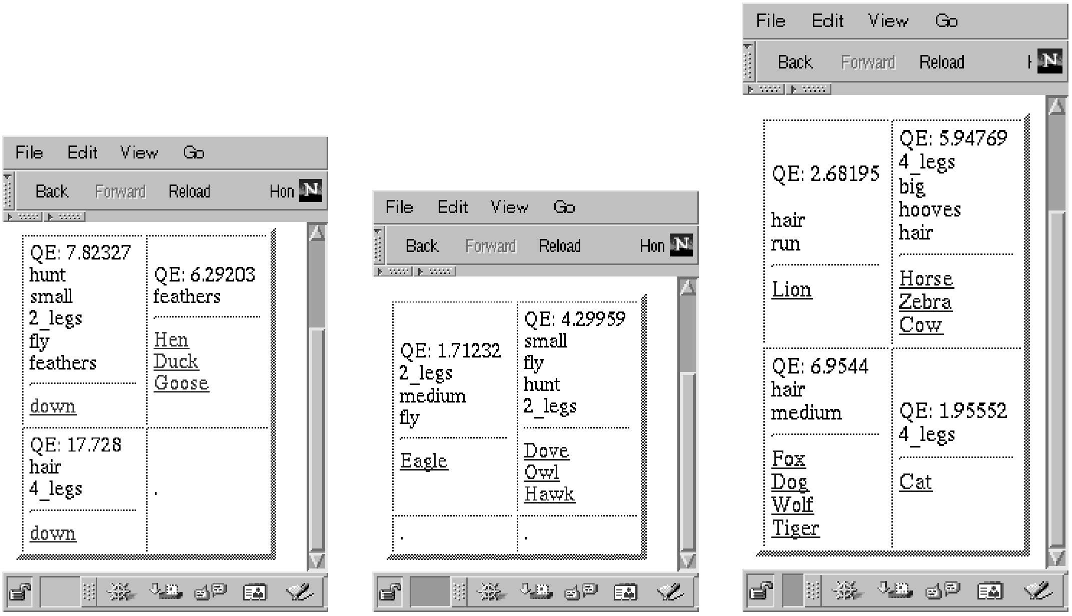
Figure 3.6: **Flat SOM:** 3 × 4 units; All animals are organized on a flat map. Mammals are located in the lower part of the map, whereas the birds are situated in the upper half.

the dove is an outlier. The unit on the right-hand represents the small hunting birds (except the dove). The *eagle* has been separated from the others due to its medium size.

On the map depicted in Figure 3.7(c), the mammals have been split up according to their size and whether they hunt or not. The hunting animals have been mapped onto the two units on the left hand-side of the map, whereas the big nonhunting animals, such as *horse*, *zebra*, and *cow* can be found on the top-right unit. The small *cat* is represented by the unit in the bottom-left corner of the map.

This example demonstrates that a hierarchical organization, as performed by the GHSOM, can offer more insight into the structure of data.

Furthermore, concerning visual data exploration, the analysis of large amounts of data is facilitated by the hierarchical representation because smaller maps are easier to survey and the danger of getting lost during browsing is reduced by



QE: 7.82327 hunt small 2_legs fly feathers down	QE: 6.29203 feathers Hen Duck Goose
QE: 17.728 hair 4_legs down	.

(a) First-Layer Map: 2 × 2 units; Rough separation into three clusters.

QE: 1.71232 2_legs medium fly Eagle	QE: 4.29959 small fly hunt 2_legs Dove Owl Hawk
.	.

(b) Second-Layer Map: 2 × 2 units; Hunting Birds

QE: 2.68195 hair run Lion	QE: 5.94769 4_legs big hooves hair Horse Zebra Cow
QE: 6.9544 hair medium Fox Dog Wolf Tiger	QE: 1.95552 4_legs Cat

(c) Second-Layer Map: 2 × 2 units; Mammals

Figure 3.7: **Hierarchy of Animals:** Nonhunting animals are situated in the top-right corner of the map. Two units have been expanded to the second layer, representing hunting birds and mammals respectively.

only showing topical subsets of the data. Just to point out this argumentation, consider a geographical map of *Europe* containing all the information that we expect a map of *Austria*, or even worse, a map of *Vienna* should contain. This hypothetical map of *Europe* will be of a size making it very difficult to find an orientation. An analogous situation occurs if a large text document collection is represented by a single map. In the next chapter we will show how a large document collection is organized hierarchically by the GHSOM.

Chapter 4

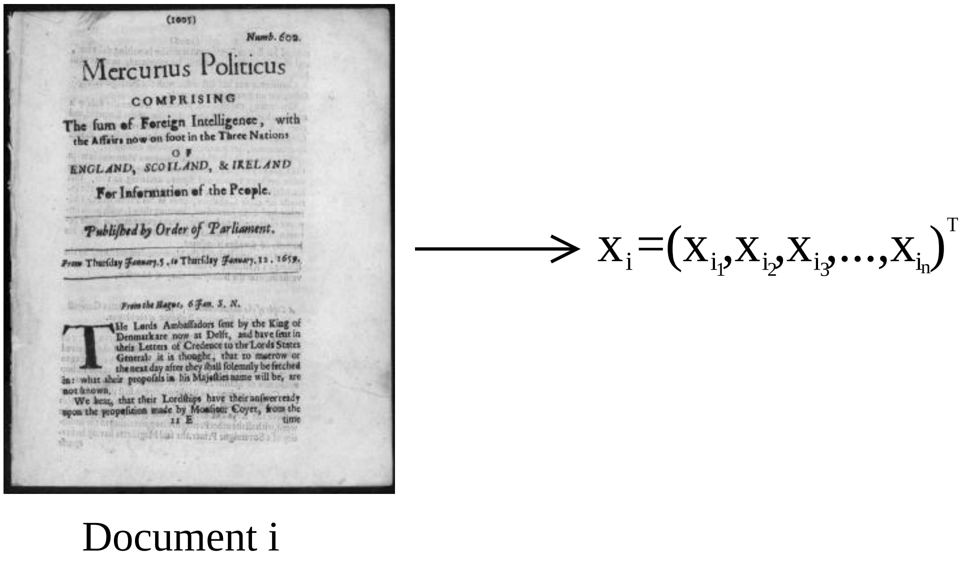
Hierarchical Clusters in Text Collections

Too little knowledge is a bad thing.
– Plato

IN this chapter we show an example from the text retrieval domain using the *Growing Hierarchical Self-Organizing Map* for text document classification and visualization. A collection of *Time Magazine* articles of the 1960's and a collection of articles from a daily Austrian newspaper have been chosen to demonstrate the organization of text documents into topical hierarchies having intuitively interpretable results.

Using text documents as input for the GHSOM implies several preprocessing steps, i.e. removing format specific tags, indexing and vector normalization, which will be explained in Section 4.1. To support the user in visually exploring the hierarchy, a method is described in Section 4.2, which automatically assigns important keywords to the units on the map to give users a hint which documents have been mapped onto these units.

Finally, both article collections and the results of the experiments using the GHSOM for document organization will be presented in Sections 4.3 and 4.4.



(1005)

Numb. 602.

Mercurius Politicus

COMPRISING

The ſum of Foreign Intelligence, with the Affairs now on foot in the Three Nations

OF

ENGLAND, SCOTLAND, & IRELAND

For Information of the People.

Publiſhed by Order of Parliament.

From Thurſday January 5, to Thurſday January 12, 1659.

From the Hague, 6 Jan. S. N.

THe Lords Ambaſſadors ſent by the King of Denmark are now at Delft, and have ſent in their Letters of Credence to the Lords States General: it is thought, that to morrow or the next day after they ſhall ſolemnly be fetched in: what their propoſals in his Majeſties name will be, are not known.

We hear, that their Lordſhips have their anſwer ready upon the propoſition made by Monſieur Coyet, from the

11 E time

Figure 4.1: **From Text to Vectors:** The vector space representation of the document is achieved by full-text indexing and weighting the importance of the index terms.

4.1 Document Representation

To be able to use text documents as input for the GHSOM a vector space representation is necessary (see Fig. 4.1[1]). The process of extracting a finite number of features to describe a set of text documents is crucial for the outcome of the SOM training process. The mapping into the two-dimensional output space can only be as good as the representation of the documents in the high-dimensional input space.

But beforehand, any format specific information has to be eliminated, such that only the pure content, i.e. the words, remains. In the case of our collection of newspaper articles the HTML tags had to be removed and so-called *entities*, used for umlauts or ligatures (e.g. `&ouml;` for ö), had to be converted into standard 7 bit ASCII characters (e.g. oe).

[1] Figure 4.1 is an image of the *Mercurius Politicus* (1659), the oldest newspaper held by the Library of Congress.

4.1.1 Full-Text Indexing and Weighting Scheme

To obtain the document vectors $x_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})^T, x_i \in \Re^n$, a list of all occurring words in the whole text collection has to be created. We will henceforth refer to this list as *template vector* where the elements of the vector are the entries of the list. The n-dimensional vector x is calculated by multiplication of the term frequency tf_{i_k} of a term k in document i with the inverse document frequency df_k (see Equation 4.1), where the document frequency is the number of documents that contain the term k.

$$x_{i_k} = tf_{i_k} \cdot \frac{1}{df_k} \tag{4.1}$$

The motivation for this weighting is that a word which occurs frequently in one document but is rare in the whole collection is significant for this very document.

The so-called tf × idf, term frequency times inverse document frequency, weighting scheme [Sal89], well known in the field of information retrieval, is a method to measure the importance of a word to represent the contents of a particular document. Pragmatically speaking, those terms are assigned large weights that appear often within a document, yet rarely within the overall document collection.

It is obvious that a large number of terms will be extracted for document representation, even in small document collections. The size of the vocabulary is also depending on the diversity of subjects.

Usually, stop word lists are used to reduce the vector dimensionality, i.e. terms such as conjunctions, articles, and pronouns, are omitted. The drawback of this method is the language dependence. Our approach was to reduce the vocabulary with respect to the document frequency of the words. Terms which occur in nearly every document and terms which are only present in a few documents are eliminated from the template vector, because both types do not contribute to a sufficient distinction between documents in the input space. Several experimental settings showed that the usual stop words mentioned above can be eliminated statistically due to their appearance frequencies. Despite this selection of the upper and lower boundary for the document frequency the vector creation process is fully automated.

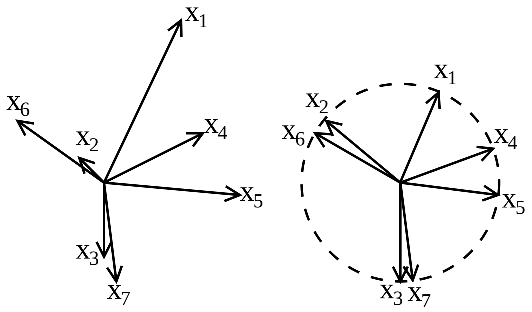


Figure 4.2: **Vector Normalization by Length:** All document vectors x_i are normalized to length 1 (a 2-dimensional graphical representation has been chosen for the sake of simplicity).

There exist several other methods for creating vector space representations of text documents like n-grams [Cav94] or *Latent Semantic Indexing* (LSI) [Dee90, Dum94]. A more mathematically motivated approach to dimensionality reduction is given in [Kas98]. The original document vectors $x_i \in \Re^n$ are multiplied by a random matrix R of dimensionality $d \times n$ resulting in vectors $y \in d$ of reduced dimensionality d. It has been shown that the data representation quality is almost as good as the original quality given a sufficiently large final dimensionality.

We chose full-text indexing and the tf × idf weighting scheme, because the labeling method described in Section 4.2 can easily be applied to the resulting maps of the GHSOM. This labeling approach assigns the important keywords for topical clusters to the units on the map. Otherwise, the significant terms would have to be extracted with further computational effort.

4.1.2 Normalization of the Input Vectors

A common practice in the text retrieval domain is to normalize the document vectors by length (see Fig. 4.2).

Using the raw tf × idf vectors is problematic due to the possibly large differences in vector length and the mapping produced by the SOM algorithm may be distorted. Experiments with a text document collection as input data using different vector normalization strategies [Dit00] exhibited that normalization by length is the most appropriate method. The quality of the mapping using vectors which were either unnormalized or normalized by attribute was by far worse.

4.2 Labeling the SOM

Not only the organization of the documents on the hierarchically structured maps is important, also the visualization has to be kept in mind. Users should be provided with a convenient interface for interactive data exploration. Currently, we use HTML tables for the representation of the maps because the hierarchical structure of independent, interlinked maps fits perfectly into the hyperlinked concept of HTML pages. Hence, the user can navigate down the hierarchy by clicking on the units of the maps, view the documents, move backwards and add bookmarks using the web-browsers capabilities.

In the text classification domain, assuming that no preclassified information (which could be used as labels) about the document collection is available, labeling the trained *Self-Organizing Map* with the determining keywords for documents mapped onto the same unit would improve the understanding of the learned organization. An automated labeling approach presented recently, namely *LabelSOM* [Rau99], can be applied to the maps of a trained *Growing Hierarchical Self-Organizing Map* to facilitate the browsing.

This method is used to assign keywords (labels) to the units of the map which are determining for the articles mapped onto these units, respectively. The labels are extracted from the vector describing the documents. As a consequence, only words which are present in the *template vector* can occur as labels and the quality of the labels is directly related to the quality of the document representation.

The weight vectors of a trained *Self-Organizing Map* can be considered as an approximation of the input vector distribution. After SOM training, the elements of a units' weight vector resemble as accurately as possible the respective vector elements of the input vectors mapped onto this particular unit and due to the topology preserving characteristic of the SOM to a certain extent those mapped onto neighboring units. Common features, i.e. vector elements, of a set of input vectors mapped onto one unit can be identified by a low variance of their values. In other words, if a majority of input vectors mapped onto the same unit show a very similar value for a particular vector element, the value of the corresponding feature in the weight vector will be highly similar as well and can therefore be considered as a common feature of the represented input vectors.

Formally, assuming S_i as the set of input vectors mapped onto unit i, $x_j \in \Re^n$ and $x_j \in S_i$ being an input vector of this set, a *quantization error* vector q_i for every unit i is calculated as the square root of the sum of the differences between the weight vector element m_{i_k} and the input vector elements x_{j_k} divided by the number $|S_i|$ of vectors mapped onto this unit for each vector element k respectively (cf. Eq. 4.2).

$$q_{i_k} = \frac{1}{|S_i|} \cdot \sqrt{\sum_{x_j \in S_i} (m_{i_k} - x_{j_k})}, \qquad k = 1 \ldots n \tag{4.2}$$

In the case of text documents, keywords describing the contents of a set of documents, are extracted. Due to the high dimensionality of the vectors representing text documents and the properties of the tf × idf weighting scheme, a large number of elements in an input vector are likely to have a value of 0 which indicates the absence of certain words in this document. The utilization of Equation 4.2 would yield a *quantization error* of 0 for such vector elements. Although the nonexistence of a word in multiple documents indicates a kind of similarity, we do not want to describe the documents in terms of commonly missing keywords. Hence, a threshold μ has to be defined to select only features with a certain relevance for describing the contents of a set of documents despite having a low *quantization error*. As described in Section 4.1.1, the higher the weight vector element's value the more important this feature is considered for representing the contents of the documents mapped onto this unit.

However, it has to be stated that the labels extracted by the *LabelSOM* algorithm are not as perfect as we would expect them to be by human selection. But due to the high dimensionality of the possibly large amount of vectors describing the documents manual selection would be nearly impossible and very time consuming. Hence, the *LabelSOM* approach is a sophisticated method providing an automated labeling which provides the user with hints on the topics organized with a SOM.

Other methods mainly focus on the coloring of *Self-Organizing Maps* to visualize the cluster structure of the data, but they offer no information on the distinct features characterizing the cluster structure. There exist several approaches like U-matrix [Ult93], projection into the CIELab color space [Kas99] or a flexible coloring method to show cluster boundaries at different levels of granularity [Him00]. The coloring techniques are based on the projection of similarity relations between neighboring units' weight vectors into a color space. Neighboring units having similar weight vectors, i.e. representing similar input vectors, are assigned other colors compared to neighboring units showing large deviations between their weight vectors.

A different approach called *Adaptive Coordinates* [Mer97] uses the movement of the weight vectors in a (virtual) output space during the SOM training process to represent the cluster structure of the data. The cluster structure is made visible by the location of the units in a two-dimensional output space, i.e. units representing similar data are grouped spatially close to each other.

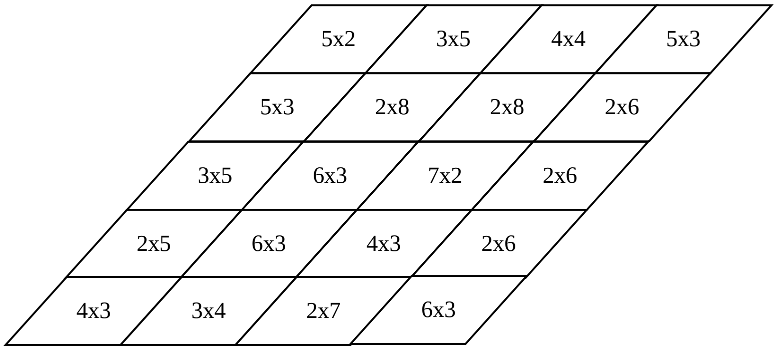
The aspect of using graphical metaphors to display the two-dimensional maps will be touched in Chapter 5.

4.3 Time Magazine

In this section, two different hierarchies of a rather small article collection consisting of 420 articles from the *TIME Magazine* are presented to demonstrate the potential of the *Growing Hierarchical Self-Organizing Map* in the field of text document classification. First, a flat hierarchy is presented with a depth of two layers and second, a deeper hierarchy with rather small maps consisting of up to four layers is shown. Furthermore, these hierarchies are compared with respect to the detected cluster structure.

4.3.1 Document Collection

The *TIME Magazine* article collection consists of 420 articles from the *TIME Magazine* of the 1960's, covering a broad range of topics from political issues to social gossip. 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 were roughly stemmed and weighted



5x2	3x5	4x4	5x3
5x3	2x8	2x8	2x6
3x5	6x3	7x2	2x6
2x5	6x3	4x3	2x6
4x3	3x4	2x7	6x3

Figure 4.3: **Shallow Time Magazine Hierarchy:** All units of the 4 × 5 first-layer map have been expanded to the second layer. The sizes of the according second-layer maps are given in the figure.

according to the tf × idf weighting scheme described in Section 4.1.1. The size of the whole collection is 2.356 kB which gives an average document size of 3.870 Bytes.

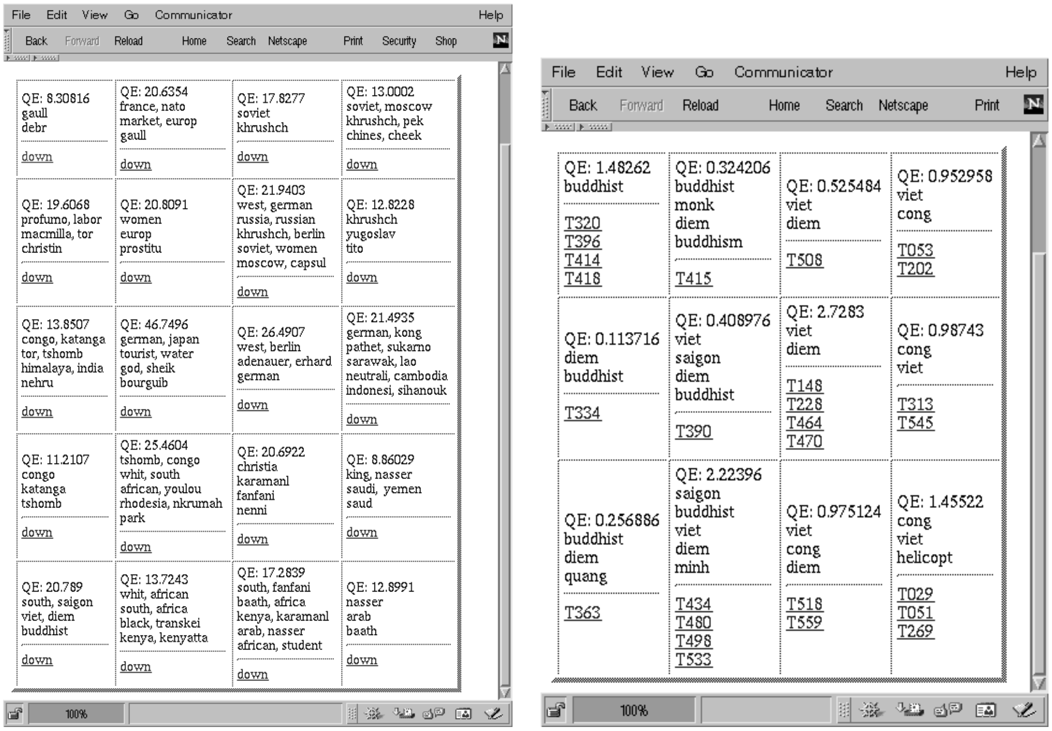
4.3.2 Shallow Hierarchy

Figure 4.3 shows a schematic representation of the hierarchy. The numbers inside the units describe the sizes of the according second-layer maps. The top-layer map evolved to a 4 × 5 map during training as depicted in Figure 4.4(a).

The topology preserving representation of the topical clusters can be found throughout the map. For example, documents covering Middle-East affairs are located in the lower right corner of the map on unit (4/4) and (4/5)[2], where the articles found on unit (4/5) deal especially with Egypt's president Gamal Abdel Nasser involved with the Syrian affairs. African topics have been mapped onto units (1/4), (2/4), (2/5), and unit (3/5). One unit covering the Vietnam War is situated in the lower left corner of the map, on unit (1/5).

If we want to take a closer look on the Vietnam topic (cf. Fig. 4.4(b)), a 4 × 3 map provides a more detailed ordering of these articles. On the units in the first and second column from the left, the articles focus on the government crack-down

[2] We refer to a unit located in column x and row y as unit (x/y), starting with (1/1) in the upper left corner.


File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop
100%
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print
100%

QE: 8.30816 gaull debr down	QE: 20.6354 france, nato market, europ gaull down	QE: 17.8277 soviet khrushch down	QE: 13.0002 soviet, moscow khrushch, pek chines, cheek down
QE: 19.6068 profumo, labor macmilla, tor christin down	QE: 20.8091 women europ prostitu down	QE: 21.9403 west, german russia, russian khrushch, berlin soviet, women moscow, capsul down	QE: 12.8228 khrushch yugoslav tito down
QE: 13.8507 congo, katanga tor, tshomb himalaya, india nehru down	QE: 46.7496 german, japan tourist, water god, sheik bourguib down	QE: 26.4907 west, berlin adenauer, erhard german down	QE: 21.4935 german, kong pathet, sukarno sarawak, lao neutrali, cambodia indonesi, sihanouk down
QE: 11.2107 congo katanga tshomb down	QE: 25.4604 tshomb, congo whit, south african, youlou rhodesia, nkrumah park down	QE: 20.6922 christia karamanl fanfani nenni down	QE: 8.86029 king, nasser saudi, yemen saud down
QE: 20.789 south, saigon viet, diem buddhist down	QE: 13.7243 whit, african south, africa black, transkei kenya, kenyatta down	QE: 17.2839 south, fanfani baath, africa kenya, karamanl arab, nasser african, student down	QE: 12.8991 nasser arab baath down

(a) First-Layer Map: 4 × 5 units; Rough representation of the main topics of the *TIME Magazine* collection.

QE: 1.48262 buddhist T320 T396 T414 T418	QE: 0.324206 buddhist monk diem buddhism T415	QE: 0.525484 viet diem T508	QE: 0.952958 viet cong T053 T202
QE: 0.113716 diem buddhist T334	QE: 0.408976 viet saigon diem buddhist T390	QE: 2.7283 viet diem T148 T228 T464 T470	QE: 0.98743 cong viet T313 T545
QE: 0.256886 buddhist diem quang T363	QE: 2.22396 saigon buddhist viet diem minh T434 T480 T498 T533	QE: 0.975124 viet cong diem T518 T559	QE: 1.45522 cong viet helicopt T029 T051 T269

(b) Second-Layer Map: 4 × 3 units; Documents covering the Vietnam War.

Figure 4.4: **Time Magazine (shallow hierarchy):** The top-layer map and the second-layer map representing the articles about the Vietnam War at greater detail.

on Buddhist monks, whereas units in the right half of the map cover the fighting and suffering during the Vietnam War.

On unit (1/4), next to unit (2/4) dealing with various articles about Africa, a cluster of documents about the Congo and the secession of its province Katanga can be found. Articles about the dispute between India and Pakistan about Kashmir are located on unit (1/3). Unit (2/1) deals with the Profumo-Keeler affair, a political scandal in Great Britain in the 1960's. Next to this unit, articles focusing on sex-related (2/2) and general gossip (2/3) can be found.

Other topics represented on this map are for example, Charles De Gaulle (1/1) and (2/1), Khrushchev (3/1) and (4/1), and Germany (3/3). While unit (1/1) deals with documents about De Gaulle and France in general, unit (2/1) covers articles with emphasis on quarrels during the Cold War between the NATO, France and other European countries about Polaris missiles and mixed crews on NATO ships.

On the second-layer map, consisting of 4 × 3 units, corresponding to unit (3/4) on the top-layer map, a detailed view on the articles about the elections in Italy is given. Furthermore, documents dealing with the aristocracy of Spain, Liechtenstein and an article about the blatant reactions on the return of Otto von Habsburg to Austria. Next to this document, two more articles about Austria can be found.

4.3.3 Deep Hierarchy

Based on the unit representing the average of all input data at layer 0, the first-layer map, initially of size 2 × 2, has grown to a size of 3 × 2 units during its training process. The units on this map have been expanded to six independent maps in the second layer (see Figure 4.5). 18 of the units in the second layer maps have further been expanded to the third layer. Three sub-topics of the third-layer maps are presented in more detail by maps on the fourth layer of the hierarchy. The units which have been expanded to the next layer are depicted as shaded circles.

One of the main topics located on the top-layer map which is shown in Figure 4.6(a) is covered by articles on African countries on unit (1/1) which can be divided into several sub-topics presented by the according second-layer map. Here we find South African issues on Apartheid, Ghana outrages, and the Egypt-Syria affairs related to Gamal Abdel Nasser. One unit has further been expanded to the third layer solely representing Nasser and Egypt related articles in more detail.

On unit (2/1) of the top layer map, the articles about the Vietnam War are located along with articles about the uproar in Ghana and some articles about the Shah in Iran (cf. Figure 4.6(b)). These topics are split up in the second layer map, where the Congo-related documents have been mapped onto unit (1/2)

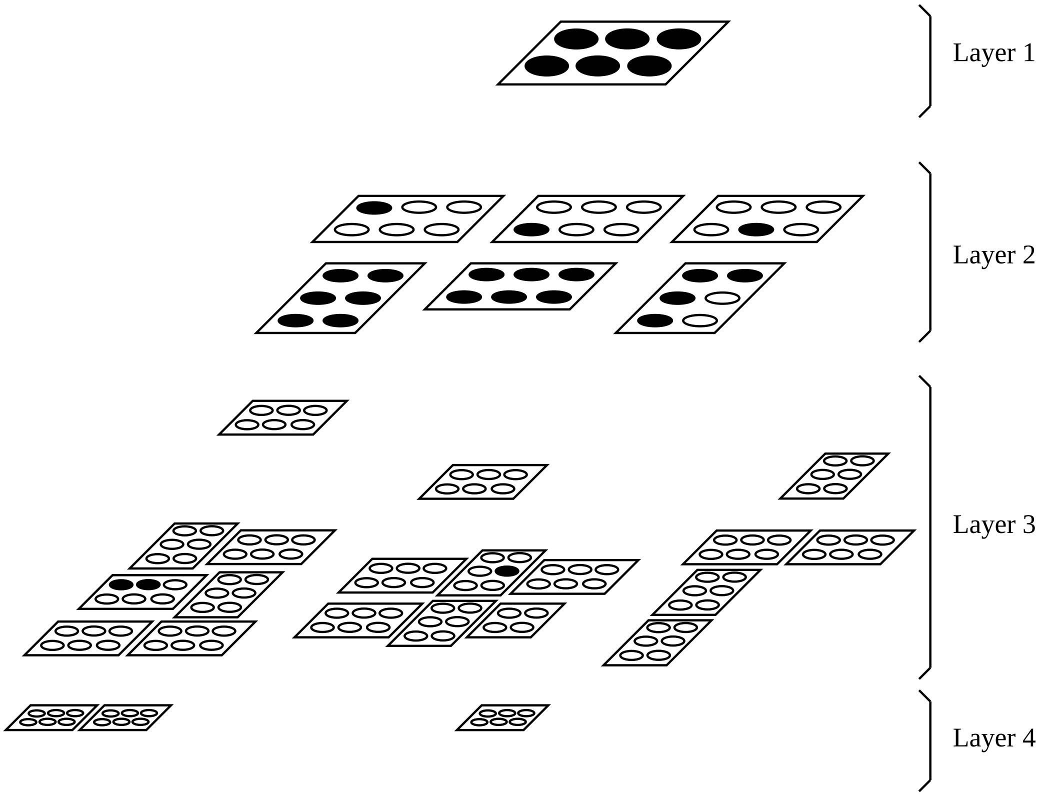
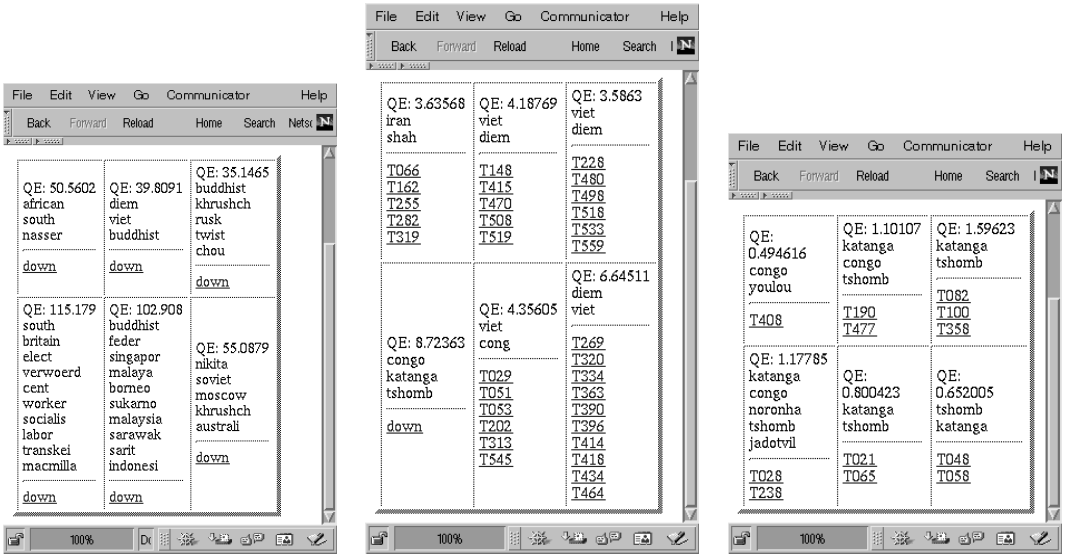


Figure 4.5: **Deep Time Magazine Hierarchy:** The branches of the GHSOM have grown to a depth of up to four layers. Units which have been expanded to the next layer are depicted as black circles.

which has been expanded to the third layer (cf. Figure 4.6(c)). The focal point of all of these articles is characterized by conflicts, fightings and outrages.

Another main subject is Nikita Khrushchev and the Soviet Union which is represented by unit (3/2). Four units of the according second-layer map have further been expanded to the third layer. Furthermore, we find various European topics like the elections in Italy, French affairs (De Gaulle), Great Britain issues, and articles about Morocco in the branch according to unit (1/2) of the top layer map. The gossip articles along with some Far East stories have been mapped onto unit (2/3).


File Edit View Go Communicator Help
Back Forward Reload Home Search Netsc
100%

QE: 50.5602 african south nasser down	QE: 39.8091 diem viet buddhist down	QE: 35.1465 buddhist khrushch rusk twist chou down
QE: 115.179 south britain elect verwoerd cent worker socialis labor transkei macmilla down	QE: 102.908 buddhist feder singapor malaya borneo sukarno malaysia sarawak sarit indonesi down	QE: 55.0879 nikita soviet moscow khrushch australi down

QE: 3.63568 iran shah T066 T162 T255 T282 T319	QE: 4.18769 viet diem T148 T415 T470 T508 T519	QE: 3.5863 viet diem T228 T480 T498 T518 T533 T559
QE: 8.72363 congo katanga tshomb down	QE: 4.35605 viet cong T029 T051 T053 T202 T313 T545	QE: 6.64511 diem viet T269 T320 T334 T363 T390 T396 T414 T418 T434 T464

QE: 0.494616 congo youlou T408	QE: 1.10107 katanga congo tshomb T190 T477	QE: 1.59623 katanga tshomb T082 T100 T358
QE: 1.17785 katanga congo noronha tshomb jadotvil T028 T238	QE: 0.800423 katanga tshomb T021 T065	QE: 0.652005 tshomb katanga T048 T058

(a) First-Layer Map: 3 × 2 units; Rough separation into three clusters.

(b) Second-Layer Map: 3×2 units; Vietnam War, Congo and Iran conflicts

(c) Third-Layer Map: 3× 2 units; The Congo conflict in detail.

Figure 4.6: **TIME Magazine (deep hierarchy):** One branch of the hierarchy (Vietnam and other violent conflicts) from the top-layer map to the third layer.

4.3.4 Comparison

The differences between both hierarchies will be highlighted by comparing some sample topical paths. In the shallow hierarchy, the documents about the Profumo-Keeler affair has been mapped onto one unit in the top-layer map together with articles about elections in Great Britain next to the units representing social gossip. In the deep hierarchy, on the other hand, the articles on the Profumo-Keeler scandal are located on unit (2/2) in the top-layer map together with the gossip articles, but separated from the other articles dealing with the British elections. These articles have been mapped onto the neighboring unit (1/2) together with articles about the Italian elections. These election-related topics have been split up on the according second-layer map of the deep hierarchy and are explained in detail on two independent third-layer maps.

The articles covering African issues are spread onto three units ((2/4), (2/5), and (3/5)) of the first-layer map of the shallow hierarchy. Articles on Ghana are located on unit (1/4) nearby unit (1/5) covering the Vietnam War. In the second experiment, the Vietnam War and Ghana issues have been mapped onto one unit, where a more detailed view is presented in the second-layer map. Here, the Ghana subject is located on a single unit which has been expanded to the second layer. The other articles concerning Africa have been organized into a different branch, namely the one according to unit (1/2) on the top-layer map.

Comparing both, the shallow and the deep hierarchy, it has to be stated that neither the first nor the second one is the "correct" hierarchical representation of the *TIME Magazine* collection. Both hierarchies show the relations between the articles depending on the desired data representation granularity which should be gained each level.

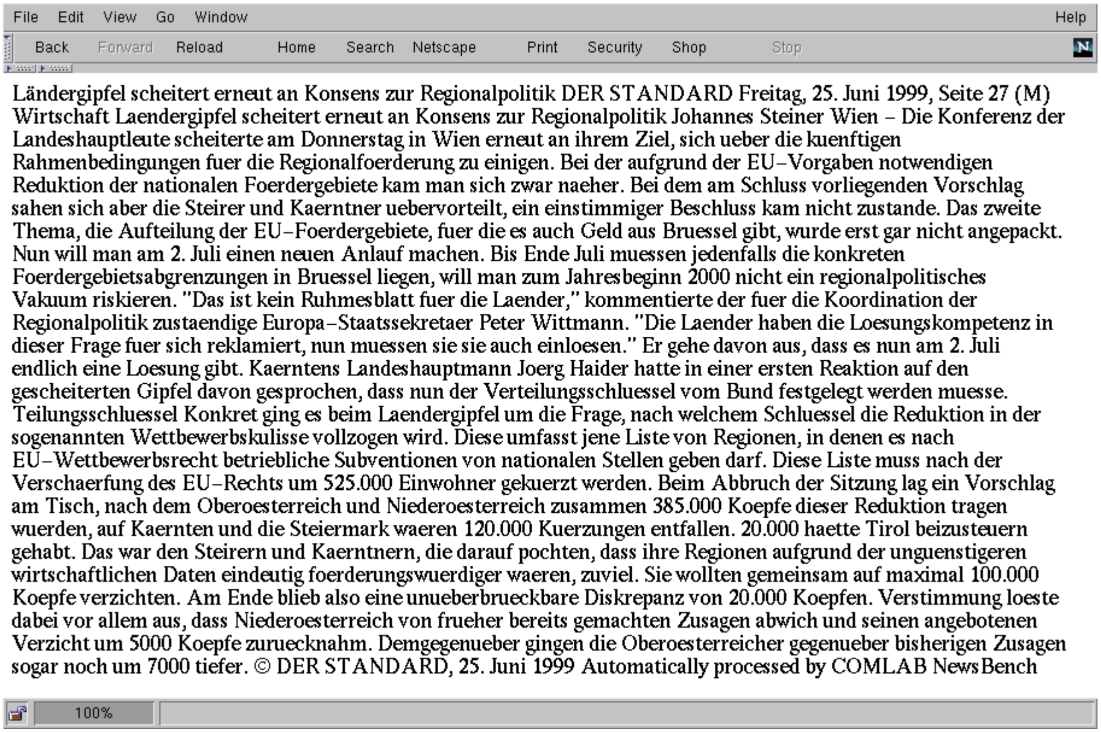
4.4 Daily Austrian Newspaper - Der Standard

4.4.1 Document Collection

In the experiments described hereafter we use an article collection of the daily Austrian newspaper *Der Standard* covering the second quarter of 1999. The articles include topics like sports, fashion, culture, politics (national and international) and others. 1104 features were automatically extracted from the 11.627 documents and again, weighted according to the *tf* × *idf* weighting scheme. The 1104 dimensions of the feature vectors already exclude terms commonly used in nearly all documents and terms which are only present in a few documents.

The size of the whole collection is 27.556 kB. Given 11.627 documents, the average length of an article is 2.426 Bytes.

A sample article about a meeting of the heads of the Austrian federal states concerning the distribution of European Union subsidies among the various regions is depicted in Figure 4.7 and examples of its important features are listed in table 4.1.



Ländergipfel scheitert erneut an Konsens zur Regionalpolitik DER STANDARD Freitag, 25. Juni 1999, Seite 27 (M) Wirtschaft Laendergipfel scheitert erneut an Konsens zur Regionalpolitik Johannes Steiner Wien – Die Konferenz der Landeshauptleute scheiterte am Donnerstag in Wien erneut an ihrem Ziel, sich ueber die kuenftigen Rahmenbedingungen fuer die Regionalfoerderung zu einigen. Bei der aufgrund der EU–Vorgaben notwendigen Reduktion der nationalen Foerdergebiete kam man sich zwar naeher. Bei dem am Schluss vorliegenden Vorschlag sahen sich aber die Steirer und Kaerntner uebervorteilt, ein einstimmiger Beschluss kam nicht zustande. Das zweite Thema, die Aufteilung der EU–Foerdergebiete, fuer die es auch Geld aus Bruessel gibt, wurde erst gar nicht angepackt. Nun will man am 2. Juli einen neuen Anlauf machen. Bis Ende Juli muessen jedenfalls die konkreten Foerdergebietsabgrenzungen in Bruessel liegen, will man zum Jahresbeginn 2000 nicht ein regionalpolitisches Vakuum riskieren. "Das ist kein Ruhmesblatt fuer die Laender," kommentierte der fuer die Koordination der Regionalpolitik zustaendige Europa–Staatssekretaer Peter Wittmann. "Die Laender haben die Loesungskompetenz in dieser Frage fuer sich reklamiert, nun muessen sie sie auch einloesen." Er gehe davon aus, dass es nun am 2. Juli endlich eine Loesung gibt. Kaerntens Landeshauptmann Joerg Haider hatte in einer ersten Reaktion auf den gescheiterten Gipfel davon gesprochen, dass nun der Verteilungsschluessel vom Bund festgelegt werden muesse. Teilungsschluessel Konkret ging es beim Laendergipfel um die Frage, nach welchem Schluessel die Reduktion in der sogenannten Wettbewerbskulisse vollzogen wird. Diese umfasst jene Liste von Regionen, in denen es nach EU–Wettbewerbsrecht betriebliche Subventionen von nationalen Stellen geben darf. Diese Liste muss nach der Verschaerfung des EU–Rechts um 525.000 Einwohner gekuerzt werden. Beim Abbruch der Sitzung lag ein Vorschlag am Tisch, nach dem Oberoesterreich und Niederoesterreich zusammen 385.000 Koepfe dieser Reduktion tragen wuerden, auf Kaernten und die Steiermark waeren 120.000 Kuerzungen entfallen. 20.000 haette Tirol beizusteuern gehabt. Das war den Steirern und Kaerntnern, die darauf pochten, dass ihre Regionen aufgrund der unguenstigeren wirtschaftlichen Daten eindeutig foerderungswuerdiger waeren, zuviel. Sie wollten gemeinsam auf maximal 100.000 Koepfe verzichten. Am Ende blieb also eine unueberbrueckbare Diskrepanz von 20.000 Koepfen. Verstimmung loeste dabei vor allem aus, dass Niederoesterreich von frueher bereits gemachten Zusagen abwich und seinen angebotenen Verzicht um 5000 Koepfe zuruecknahm. Demgegenueber gingen die Oberoesterreicher gegenueber bisherigen Zusagen sogar noch um 7000 tiefer. © DER STANDARD, 25. Juni 1999 Automatically processed by COMLAB NewsBench

Figure 4.7: **Newspaper Article:** Sample document about Austrian and European Union issues.

4.4.2 Walking through a Newspaper Archive

Since it is impossible to present the complete topic hierarchy of three months of daily newspaper articles, we will concentrate on some sample topical sections. Furthermore, for the sake of readability, the layout of some HTML tables has been edited in order to fit the screenshots to the pages of this thesis. On some maps, where a multitude of documents have been mapped onto one unit (e.g. Fig. 4.11), the list of the articles has been replaced by the number of documents which have been mapped onto the respective unit.

The top-level map of the GHSOM, which evolved to the size of 3×4 units during training by adding two rows and one column, is depicted in Figure 4.8. This map represents the organization of the main subject areas such as the war on the Balkan on unit (1/1), Austrian internal politics (2/1) and (2/2), economy (3/1) and the European Union on unit (3/2). Furthermore, we can find computers

keyword	**tf × idf values**
EU	7.78
Lower Austria	7.05
Federal States	6.59
Brussels	6.18

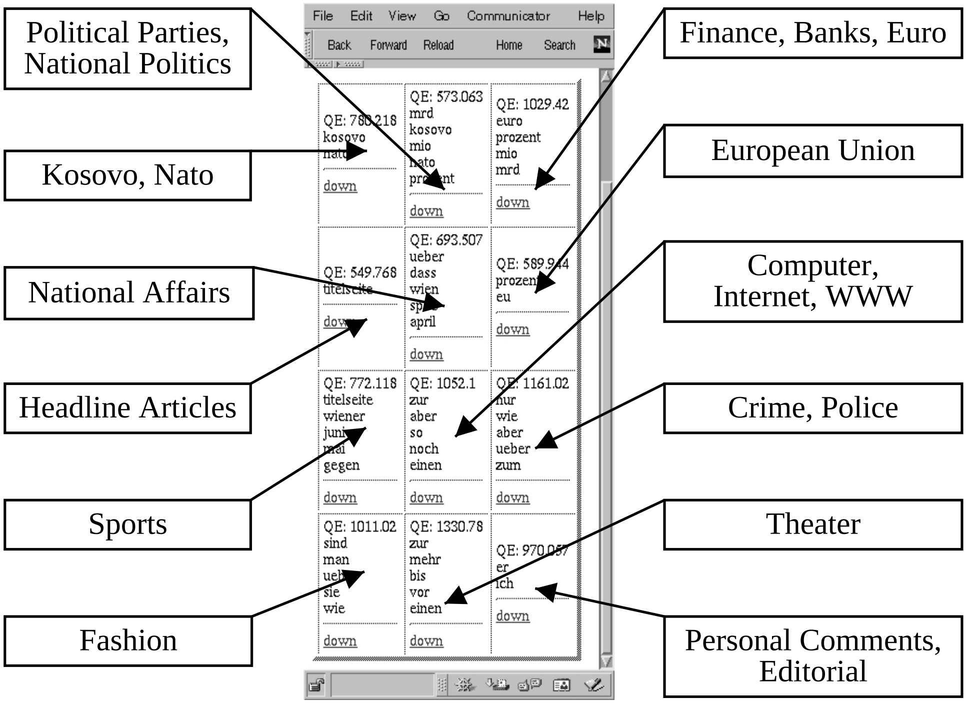
Table 4.1: **Important Features:** Some keywords (translated to English) having relatively high *tf* × *idf* values (before vector normalization) of the article presented in Fig. 4.7.

and Internet on unit (2/3) and articles on crime and police on unit (3/3). Cultural topics are located on the three units in the lower left corner, representing sports, theater, and fashion, respectively.

The labels which have been generated with the *LabelSOM* algorithm are not in any case suitable, especially for the topics present in the lower half of the map. The cause for this inaccuracy is the large diversity of articles located on a single unit due to the rough organization at the first layer. Particularly the cultural topic consists of a variety of different subtopics which cannot be characterized by specific words. The labeling of the units in the upper half of the map worked because of several keywords being predominantly present, e.g. *kosovo*, *NATO* on unit (1/1). However, different labeling method for the top-layer map has to be thought of.

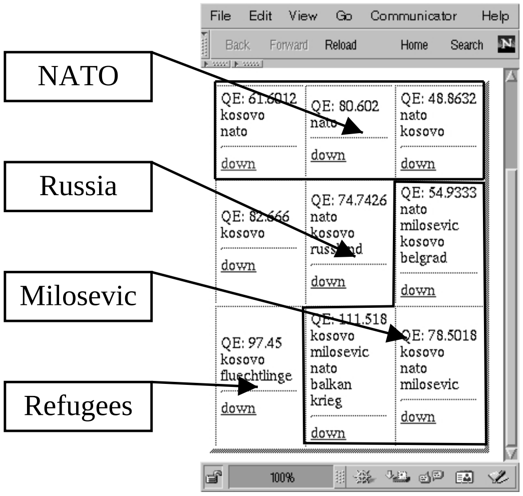
If we want to take a closer look at the Balkan War subject, we discover a more detailed map of size 3 × 3 in the next layer (see Fig. 4.9), where every unit represents a specific sub-topic of this subject. The units in the top row deal with the general political and social situation in Kosovo while the main focus of the documents mapped onto the three units (3/2), (2/3) and (3/3) in the lower right corner is Slobodan Milosevic. On unit (1/3) articles about the situation of the refugees can be found, whereas unit (2/2) concentrates on the Russian involvement in the Kosovo War.

If this Russian involvement is our major interest, again, a more detailed map



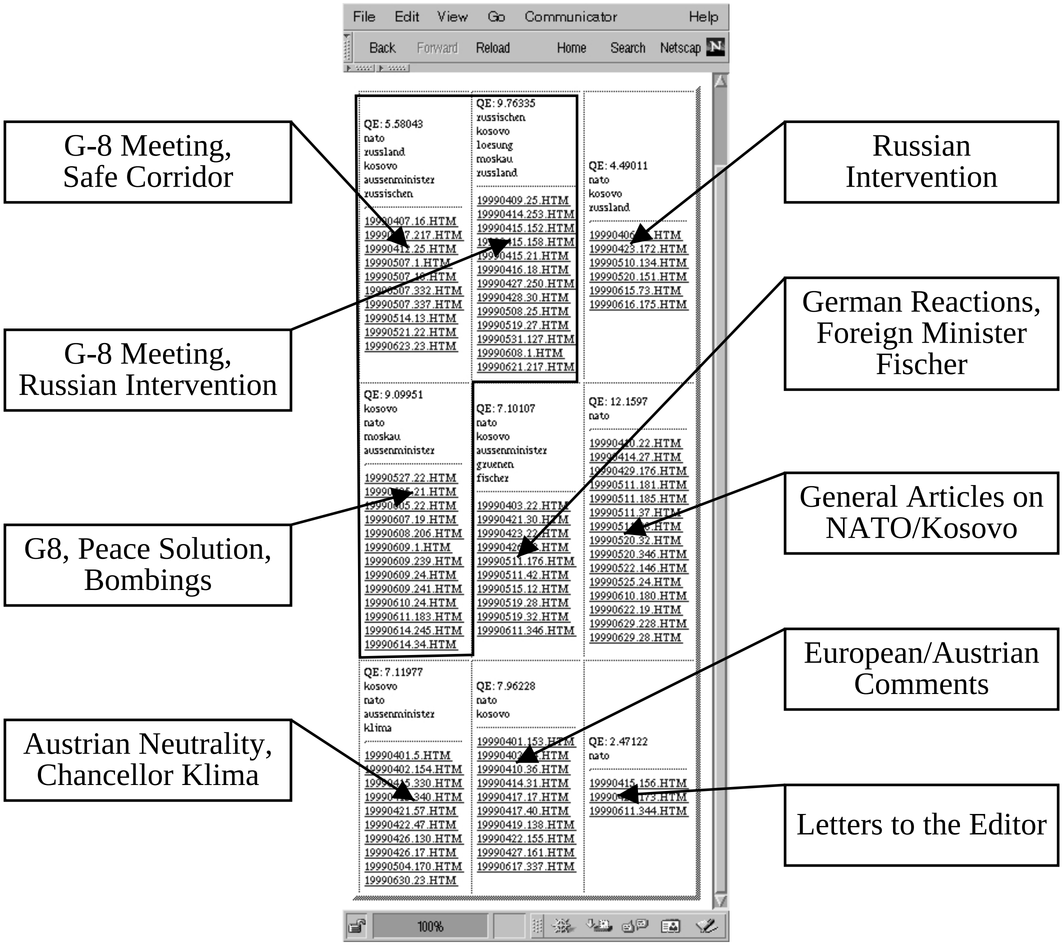
QE: 780.218 kosovo nato down	QE: 573.063 mrd kosovo mio nato prozent down	QE: 1029.42 euro prozent mio mrd down
QE: 549.768 titelseite down	QE: 693.507 ueber dass wien sport april down	QE: 589.944 prozent eu down
QE: 772.118 titelseite wiener juni mai gegen down	QE: 1052.1 zur aber so noch einen down	QE: 1161.02 nur wie aber ueber zum down
QE: 1011.02 sind man ueber sie wie down	QE: 1330.78 zur mehr bis vor einen down	QE: 970.057 er ich down

Figure 4.8: **Top-Layer Map:** 3 × 4 units; Organization of main subject areas


NATO
Russia
Milosevic
Refugees
File Edit View Go Communicator Help
Back Forward Reload Home Search
100%

QE: 61.6012 kosovo nato down	QE: 80.602 nato down	QE: 48.8632 nato kosovo down
QE: 82.666 kosovo down	QE: 74.7426 nato kosovo russland down	QE: 54.9333 nato milosevic kosovo belgrad down
QE: 97.45 kosovo fluechtlinge down	QE: 111.518 kosovo milosevic nato balkan krieg down	QE: 78.5018 kosovo nato milosevic down

Figure 4.9: **Second-Layer Map:** 3 × 3 units; corresponds to unit (2/1) on the top-layer map; Balkan War



QE: 5.58043 nato russland kosovo aussenminister russischen 19990407.16.HTM 1999[illegible]7.217.HTM 19990412.25.HTM 19990507.1.HTM 19990507.[illegible].HTM 19990507.332.HTM 19990507.337.HTM 19990514.13.HTM 19990521.22.HTM 19990623.23.HTM	QE: 9.76335 russischen kosovo loesung moskau russland 19990409.25.HTM 19990414.253.HTM 19990415.152.HTM 19990415.158.HTM 19990415.21.HTM 19990416.18.HTM 19990427.250.HTM 19990428.30.HTM 19990508.25.HTM 19990519.27.HTM 19990531.127.HTM 19990608.1.HTM 19990621.217.HTM	QE: 4.49011 nato kosovo russland 19990406.[illegible].HTM 19990423.172.HTM 19990510.134.HTM 19990520.151.HTM 19990615.73.HTM 19990616.175.HTM
QE: 9.09951 kosovo nato moskau aussenminister 19990527.22.HTM 19990[illegible]5.21.HTM 19990[illegible]05.22.HTM 19990607.19.HTM 19990608.206.HTM 19990609.1.HTM 19990609.239.HTM 19990609.24.HTM 19990609.241.HTM 19990610.24.HTM 19990611.183.HTM 19990614.245.HTM 19990614.34.HTM	QE: 7.10107 nato kosovo aussenminister gruenen fischer 19990403.22.HTM 19990421.30.HTM 19990423.22.HTM 1999042[illegible].HTM 19990511.176.HTM 19990511.42.HTM 19990515.12.HTM 19990519.28.HTM 19990519.32.HTM 19990611.346.HTM	QE: 12.1597 nato 19990410.22.HTM 19990414.27.HTM 19990429.176.HTM 19990511.181.HTM 19990511.185.HTM 19990511.37.HTM 199905[illegible].HTM 19990520.32.HTM 19990520.346.HTM 19990522.146.HTM 19990525.24.HTM 19990610.180.HTM 19990622.19.HTM 19990629.228.HTM 19990629.28.HTM
QE: 7.11977 kosovo nato aussenminister klima 19990401.5.HTM 19990402.154.HTM 19990415.330.HTM 19990[illegible].340.HTM 19990421.57.HTM 19990422.47.HTM 19990426.130.HTM 19990426.17.HTM 19990504.170.HTM 19990630.23.HTM	QE: 7.96228 nato kosovo 19990401.153.HTM 1999040[illegible].HTM 19990410.36.HTM 19990414.31.HTM 19990417.17.HTM 19990417.40.HTM 19990419.138.HTM 19990422.155.HTM 19990427.161.HTM 19990617.337.HTM	QE: 2.47122 nato 19990415.156.HTM 19990[illegible].173.HTM 19990611.344.HTM

G-8 Meeting,
Safe Corridor

G-8 Meeting,
Russian Intervention

G8, Peace Solution,
Bombings

Austrian Neutrality,
Chancellor Klima

Russian
Intervention

German Reactions,
Foreign Minister
Fischer

General Articles on
NATO/Kosovo

European/Austrian
Comments

Letters to the Editor

Figure 4.10: **Third-Layer Map:** 3×3 units; Russian involvement in the Balkan War

(see Fig. 4.10) in the third layer exhibits the most granular view on this sub-topic[3]. The dominant topic of this map is the Russian embroilment in the Kosovo War. The three units in the upper left corner cover *G-8* foreign minister meetings, where the documents on unit (1/1) are especially about the establishment of a safe corridor to the Kosovo. Articles located on units (1/2) and (2/1) discuss other Kosovo-related aspects of these meetings. On unit (2/2) and (1/3) we can find articles about reactions of the German foreign minister Fischer and Austrian chancellor Klima respectively. European comments have been mapped onto unit (2/3) and letters to the editor onto unit (3/3).

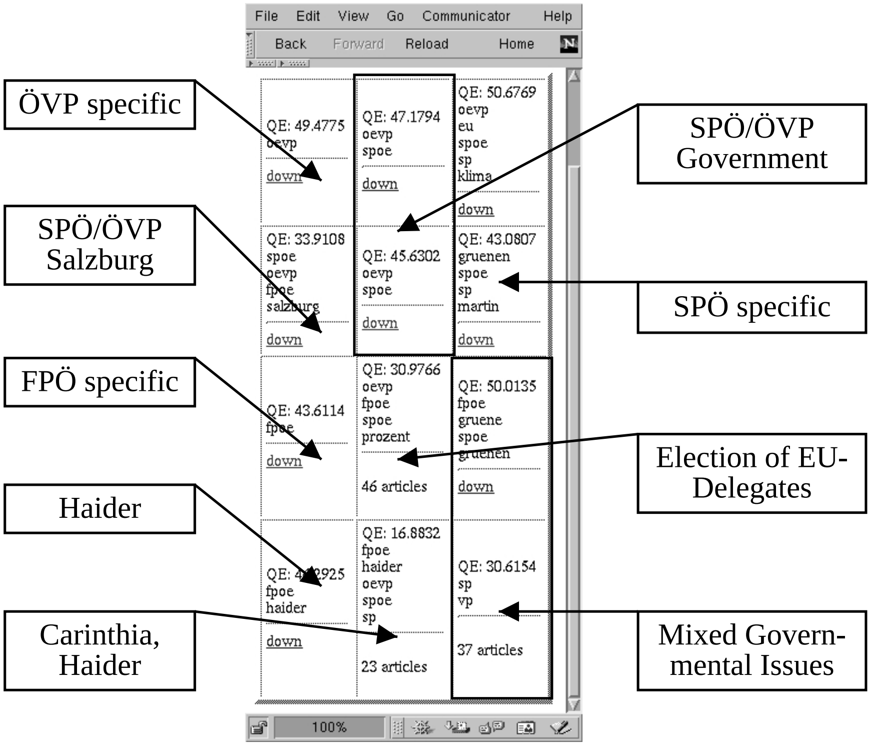
Another main subject found on unit (2/1) of the top-level map (see Fig. 4.8) can be viewed in greater detail in the second-layer map depicted in Figure 4.11. This particular map contains documents dealing with the political situation in Austria. In particular, Austrian political parties are covered and ordered according to the various sub-topics.

Articles dealing with the governing coalition (Social Democrats (SPÖ) and conservative People's Party (ÖVP) at that time) are represented by units in the upper half of the map, where documents with emphasis on the ÖVP are located on unit (1/1), articles dealing with SPÖ subjects on unit (3/2) and articles about both parties in between. Articles about the opposing Freedom Party (FPÖ) can be found on units (1/3), (1/4) and (2/4), whereby the latter two focus on the leader J. Haider. Here, the possible unbalancedness of the GHSOM, discussed in Chapter 3, can be observed. Units (2/3) and (2/4) have not been expanded into the next layer due to the rather high similarity of the documents' contents.

Other Austrian issues are organized on the second-level map shown in Figure 4.12, where clusters covering building projects and the according scandals (1/1), various regional issues (1/2) and smaller political topics distributed over the rest of the map can be identified. Articles about Austrian media have been mapped onto unit (4/3) and reports with an topical overlap of politics and media onto unit (3/3).

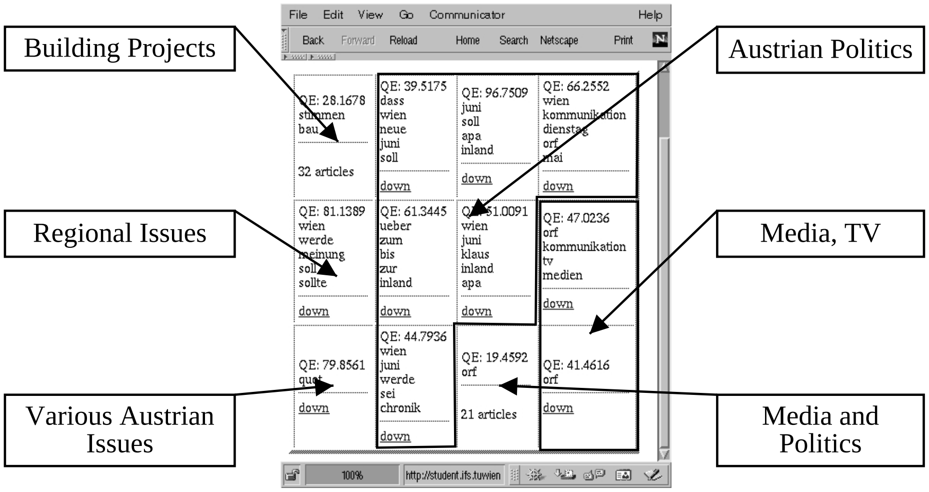
The organization of articles covering economic sub-topics (unit (3/1) on the top-layer map) can be explored in a more granular view on the according second-level map consisting of 5 × 2 units (see Fig. 4.13). Articles concerning the bank

[3] The most granular view with regard to the settings of τ_2 as described in Section 3.3.



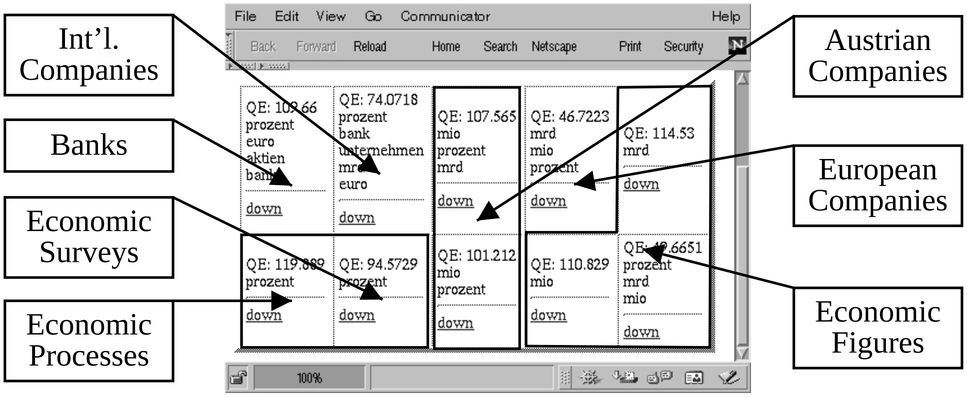
QE: 49.4775 oevp down	QE: 47.1794 oevp spoe down	QE: 50.6769 oevp eu spoe sp klima down
QE: 33.9108 spoe oevp fpoe salzburg down	QE: 45.6302 oevp spoe down	QE: 43.0807 gruenen spoe sp martin down
QE: 43.6114 fpoe down	QE: 30.9766 oevp fpoe spoe prozent 46 articles	QE: 50.0135 fpoe gruene spoe gruenen down
QE: 4[illegible]2925 fpoe haider down	QE: 16.8832 fpoe haider oevp spoe sp 23 articles	QE: 30.6154 sp vp 37 articles

Figure 4.11: **Second-Layer Map:** 3 × 4 units; corresponds to unit (2/1) on the top-layer map; Austrian politics


File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print
Building Projects
Regional Issues
Various Austrian Issues
Austrian Politics
Media, TV
Media and Politics
100%
http://student.ifs.tuwien

QE: 28.1678 stimmen bau 32 articles	QE: 39.5175 dass wien neue juni soll down	QE: 96.7509 juni soll apa inland down	QE: 66.2552 wien kommunikation dienstag orf mai down
QE: 81.1389 wien werde meinung soll sollte down	QE: 61.3445 ueber zum bis zur inland down	QE: 51.0091 wien juni klaus inland apa down	QE: 47.0236 orf kommunikation tv medien down
QE: 79.8561 quot down	QE: 44.7936 wien juni werde sei chronik down	QE: 19.4592 orf 21 articles	QE: 41.4616 orf down

Figure 4.12: **Second-Layer Map:** 4 × 3 units; corresponds to unit (2/2) on the top-layer map; Austrian affairs



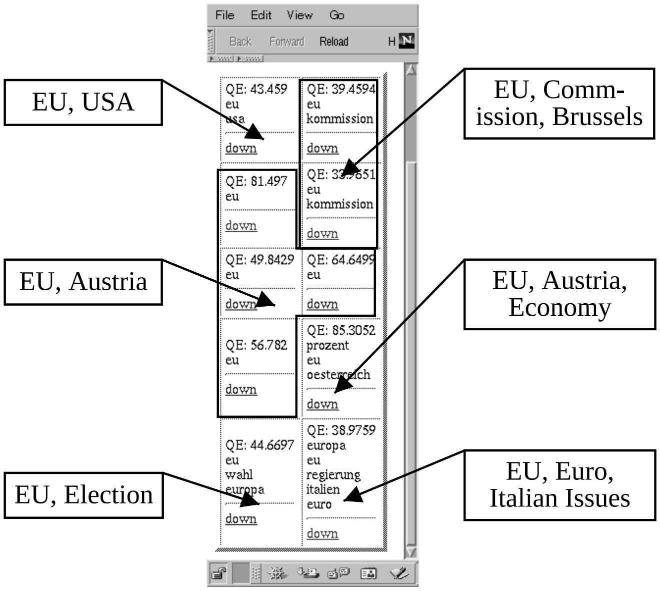
QE: 102.66 prozent euro aktien bank down	QE: 74.0718 prozent bank unternehmen mrd euro down	QE: 107.565 mio prozent mrd down	QE: 46.7223 mrd mio prozent down	QE: 114.53 mrd down
QE: 119.889 prozent down	QE: 94.5729 prozent down	QE: 101.212 mio prozent down	QE: 110.829 mio down	QE: 49.6651 prozent mrd mio down

Figure 4.13: **Second-Layer Map:** 5 × 2 units; corresponds to unit (3/1) on the top-layer map; Economy

domain (1/1), international companies (2/1), Austrian companies (3/1) and (3/2) and European companies (4/1) are covered by this map. Units (5/1), (4/2) and (5/2) in the lower right corner deal with reports containing lots of economic figures, whereas in the lower left corner articles about economic processes and surveys are located.

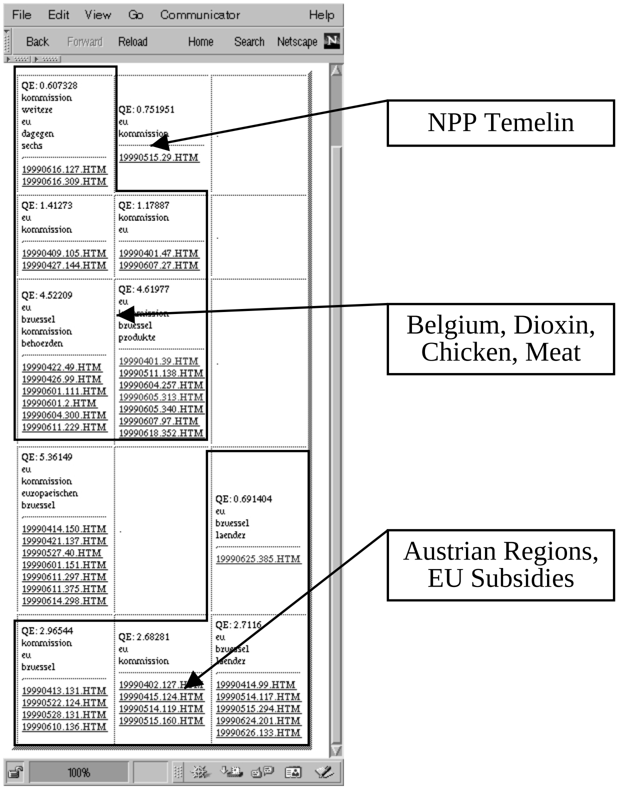
The European Union topic mapped onto unit (3/2) in the first-layer map has been expanded to a second-level map depicted in Figure 4.14. It has been organized into sub-topics like EU – USA connections on unit (1/1), the European Commission on units (2/1) and (2/2) or a large cluster of articles about the relationship between Austria and the EU in the middle of the map. Unit (2/4) concentrates on economic aspects of the Austrian membership in the European Union and unit (2/5) represents documents covering Italian affairs.

Figure 4.15 exhibits a view on the ordering of documents mapped onto unit (2/2) in the corresponding second-layer map described before. In the upper half we find articles about the scandal in Belgium where chicken meat was contaminated by large amounts of dioxin and discussions about a ban on meat exports. In the bottom row of the map documents concentrating on European Union subsidies and discussions about the distribution of these fundings between different Austrian regions are located. The sample article depicted in Figure 4.7 can be found on unit (3/3). In between, on unit (2/3), reports on genetically modified



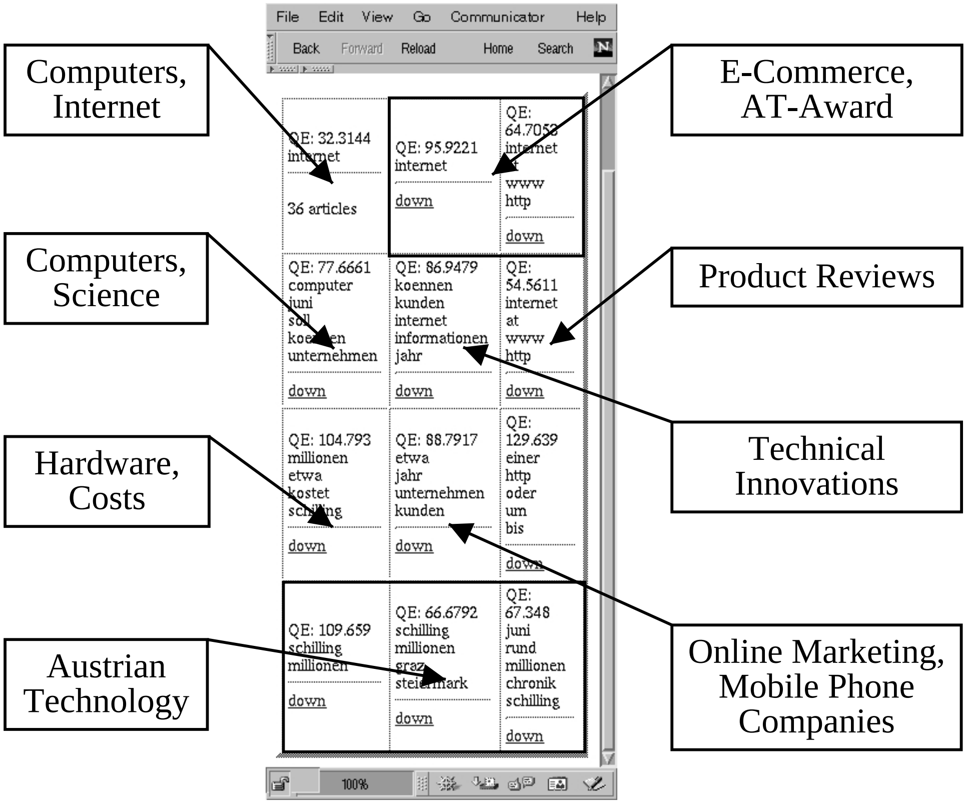
QE: 43.459 eu usa down	QE: 39.4594 eu kommission down
QE: 81.497 eu down	QE: 33.9651 eu kommission down
QE: 49.8429 eu down	QE: 64.6499 eu down
QE: 56.782 eu down	QE: 85.3052 prozent eu oesterreich down
QE: 44.6697 eu wahl europa down	QE: 38.9759 europa eu regierung italien euro down

Figure 4.14: **Second-Layer Map:** 2 × 5 units; European Union topics



QE: 0.607328 kommission weitere eu dagegen sechs 19990616.127.HTM 19990616.309.HTM	QE: 0.751951 eu kommission 19990515.29.HTM	.
QE: 1.41273 eu kommission 19990409.105.HTM 19990427.144.HTM	QE: 1.17887 kommission eu 19990401.47.HTM 19990607.27.HTM	.
QE: 4.52209 eu bruessel kommission behoerden 19990422.49.HTM 19990426.99.HTM 19990601.111.HTM 19990601.2.HTM 19990604.300.HTM 19990611.229.HTM	QE: 4.61977 eu kommission bruessel produkte 19990401.39.HTM 19990511.138.HTM 19990604.257.HTM 19990605.313.HTM 19990605.340.HTM 19990607.97.HTM 19990618.352.HTM	.
QE: 5.36149 eu kommission europaeischen bruessel 19990414.150.HTM 19990421.137.HTM 19990527.40.HTM 19990601.151.HTM 19990611.297.HTM 19990611.375.HTM 19990614.298.HTM	.	QE: 0.691404 eu bruessel laender 19990625.385.HTM
QE: 2.96544 kommission eu bruessel 19990413.131.HTM 19990522.124.HTM 19990528.131.HTM 19990610.136.HTM	QE: 2.68281 eu kommission 19990402.127.HTM 19990415.124.HTM 19990514.119.HTM 19990515.160.HTM	QE: 2.7116 eu bruessel laender 19990414.99.HTM 19990514.117.HTM 19990515.294.HTM 19990624.201.HTM 19990626.133.HTM

Figure 4.15: **Third-Layer Map:** 3 × 4 units; Agricultural and regional topics



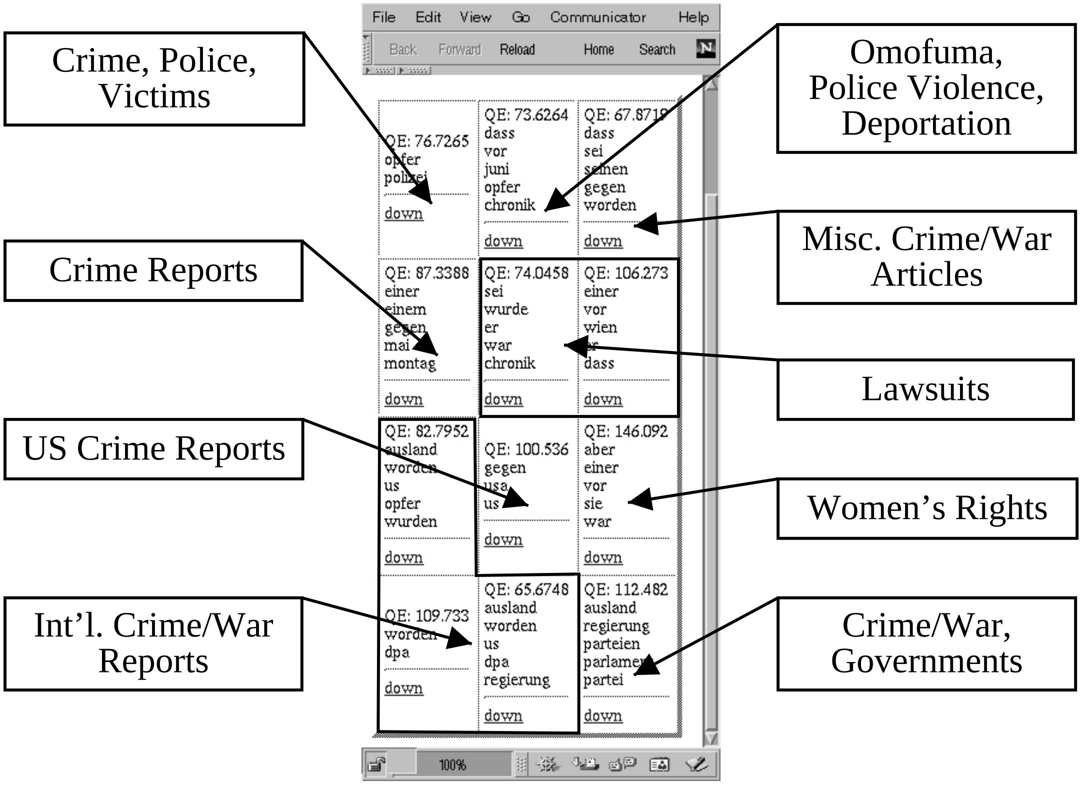
QE: 32.3144 internet 36 articles	QE: 95.9221 internet down	QE: 64.7053 internet at www http down
QE: 77.6661 computer juni soll koennen unternehmen down	QE: 86.9479 koennen kunden internet informationen jahr down	QE: 54.5611 internet at www http down
QE: 104.793 millionen etwa kostet schilling down	QE: 88.7917 etwa jahr unternehmen kunden down	QE: 129.639 einer http oder um bis down
QE: 109.659 schilling millionen down	QE: 66.6792 schilling millionen graz steiermark down	QE: 67.348 juni rund millionen chronik schilling down

Figure 4.16: **Second-Layer Map:** 3×4 units; corresponds to unit (2/3) on the top-layer map; Computers and Internet

food and arguments about sale stops of Belgian goods in Austrian supermarkets can be found.

Information technology can be identified on unit (2/3) in the top-layer map as another major subject and Figure 4.16 depicts the according second-layer map. Starting with articles about Internet and computers on unit (1/1), the units (2/1) and (3/1) cover E-Commerce and the “at award”[4]. Unit (1/2) represents scientific aspects and unit (2/2) concentrates on reports on technological innovations, whereas product reviews can be found on unit (3/2) and articles about mobile phone companies and service providers are located on unit (2/3). Documents concerning Austria and technology have been mapped onto the units in the bottom row of the map.

[4] Internet award presented by the newspaper publishing company.

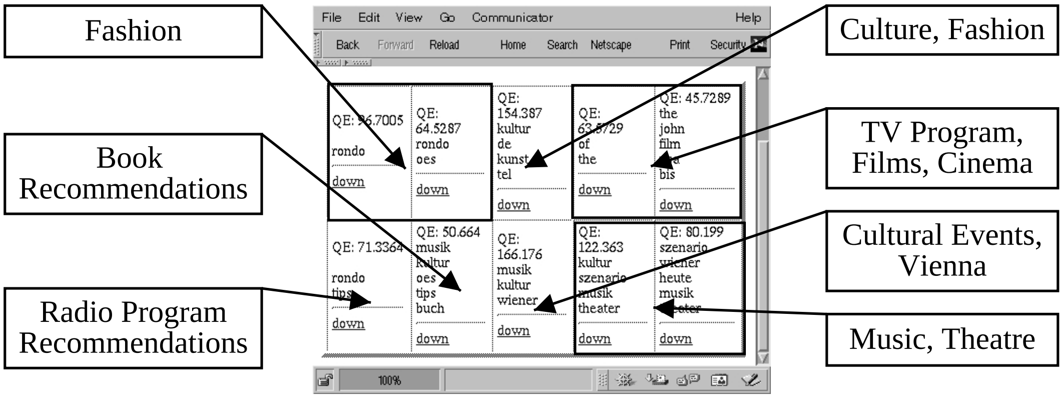


QE: 76.7265 opfer polizei down	QE: 73.6264 dass vor juni opfer chronik down	QE: 67.8719 dass sei seinen gegen worden down
QE: 87.3388 einer einem gegen mai montag down	QE: 74.0458 sei wurde er war chronik down	QE: 106.273 einer vor wien er dass down
QE: 82.7952 ausland worden us opfer wurden down	QE: 100.536 gegen usa us down	QE: 146.092 aber einer vor sie war down
QE: 109.733 worden dpa down	QE: 65.6748 ausland worden us dpa regierung down	QE: 112.482 ausland regierung parteien parlament partei down

Figure 4.17: **Second-Layer Map:** 3 × 4 units; corresponds to unit (3/3) on the top-layer map; Crime and police reports

Articles focusing on wars, crime and police which have been mapped onto unit (3/3) in the first-layer map are organized on a second-layer map of size 3×4 presented in Figure 4.17. Diverse crime reports are distributed over the units in the leftmost row, separated by the origin of the crime. Units (1/1) and (2/2) deal especially with reports about police actions, where the documents located on unit (2/2) cover the Austrian *Omofuma* scandal in which a prisoner died due to suffocation during deportation to his home country. Articles on lawsuits can be found on units (2/2) and (3/2), whereas documents about women's rights and issues are represented by unit (3/3).

Articles concerning culture and cultural events (see unit (1/4) on the first-layer map) can be explored in higher granularity on the map in the second layer shown in Figure 4.18. A weekend addendum covering fashion has been mapped onto two units in the upper left corner of the map. A mixture of articles on


Fashion
Book Recommendations
Radio Program Recommendations
Culture, Fashion
TV Program, Films, Cinema
Cultural Events, Vienna
Music, Theatre
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security
100%

QE: 96.7005 rondo down	QE: 64.5287 rondo oes down	QE: 154.387 kultur de kunst tel down	QE: 63.5729 of the down	QE: 45.7289 the john film [illegible] bis down
QE: 71.3364 rondo tips down	QE: 50.664 musik kultur oes tips buch down	QE: 166.176 musik kultur wiener down	QE: 122.363 kultur szenario musik theater down	QE: 80.199 szenario wiener heute musik theater down

Figure 4.18: **Second-Layer Map:** 5 × 2 units; corresponds to unit (1/4) on the top-layer map; Culture

culture and fashion can be found on unit (3/1). Units (1/2) and (2/2) represent recommendations for radio programs and books respectively. Nearby, articles on cultural events (3/2) and music/theater, units (3/2) and (4/2), are located. Finally, TV programs, information on TV films and cinema can be found in the top right corner of the map.

Some problems arose from the relatively short document description used in the experiment presented herein. The selection of the vocabulary describing the documents had to be rather restrictive due to the current memory-wasting implementation of the GHSOM. In some cases, articles were misclassified because of missing descriptive terms which were omitted by virtue of their insufficient document frequency. However, this overall impression of this result obtained with the *Growing Hierarchical Self-Organizing Map*, are very promising.

4.5 Discussion

In this chapter we presented a novel method, how a neural network architecture can be used for automatic organization of text documents.

The result presented herein shows the dynamic adaptation of the *Growing Hierarchical Self-Organizing Map* according to the structure of the *TIME Magazine* collection and a newspaper article collection. A rough representation of the articles can be found on the rather small map in the first layer, where only the major subjects can be identified. Being interested in a specific topic, one can browse downwards through the hierarchy whereby the granularity of data representation increases layer by layer, i.e. the according subject is split into multiple sub-topics, these are split again and so on, until the actual documents can be accessed through a map displaying them at the highest resolution determined by the training parameters.

Several other document collections the GHSOM has been used with, include the *CIA World Factbook* describing the countries of the world, a complete year of the *Scientific American*, and legal text corpora which have been classified within the *Konterm III* project[Sch99, Sch00].

It has to be noted that the GHSOM as well as the *Self-Organizing Map* is not limited to organizing text collections. Some preliminary experiments with images described either by color histograms or texture information produced promising results. A more sophisticated example for content-based image retrieval using the self-organizing map can be found in [Laa99].

Chapter 5

Conclusion and Future Work

There was a point to this story,
but it has temporarily escaped
the chronicler's mind.
– Douglas Adams' "So long, and thank's for all the fish"

IN this thesis we presented a novel neural network architecture with an unsupervised learning rule. This highly adaptive architecture, namely *Growing Hierarchical Self-Organizing Map* (GHSOM), is a combination of several neural network models described in Chapter 2. The *Self-Organizing Map* [Koh95] is the foundation of the work presented in this thesis. The SOM has proven to be a highly effective tool for cluster analysis and visualization of high-dimensional data. Due to the fixed size of the neuron grid, which has to be defined prior to the training process, several problems arise when the SOM is used for large data sets.

First, the size of the *Self-Organizing Map* is determining the resolution at which the data will be displayed. Therefore, a large number of input data requires a large map which is not handy for exploratory data analysis, because it is difficult to find an orientation within and keep an overview of the data. Second, with increasing size of the maps the computational load also increases which can be an obstacle to using SOMs in real-world applications.

The drawback of the predefined size of the network is eliminated by the *Growing Grid* architecture [Fri95]. At the beginning of the training process the map is of rather small size. During training rows or columns of units are inserted into the map where the input space is likely to be underrepresented. This growth process can be terminated by any stopping criterion suitable for the data set or another application specific requirement. Hence, the map size is determined dynamically during the training process. This feature still does not solve the problem of large map sizes when having a large set of input data, because after the growth process the resulting map will still be large.

The *Hierarchical Feature Map* [Mii90] follows a different approach. It consists of several layers of independent *Self-Organizing Maps* which are trained sequentially from the top layer down to the bottom layer. For every unit in a layer, a map in the next layer is added. This balanced tree of *Self-Organizing Maps*, which are predefined in size, has a fixed number of layers. The single first-layer map is trained with the complete input data set until a stable state is reached. Then, the maps in the next layer are trained, but only with the respective subset of the data which has been mapped onto the according unit in the upper layer. The outcome of this architecture resembles the hierarchical structure of the input data and the partitioning of the data onto several maps provides a convenient interface for exploratory data analysis because of the small maps. The difficulty of appropriately defining the size of this neural network model is even harder compared to the SOM, because the depth of the hierarchy has to be defined in addition to the size of the various maps on each layer. Furthermore, the possible ununiformity of the input data distribution must not fit adequately into a balanced structure.

Our GHSOM approach (detailed in Chapter 3) combines the advantages of the above mentioned neural networks and eliminates some drawbacks by using growing *Self-Organizing Maps* in a dynamically growing hierarchical structure. Hence, each map grows until the desired quality of data representation is reached (starting with a rather rough representation at the top layer) and maps are added to the hierarchy dynamically during training until the (usually high) data representation quality criterion at the lowest level is met. This architecture adapts itself according to the requirements of the input data in contrast to force the data into a predefined structure.

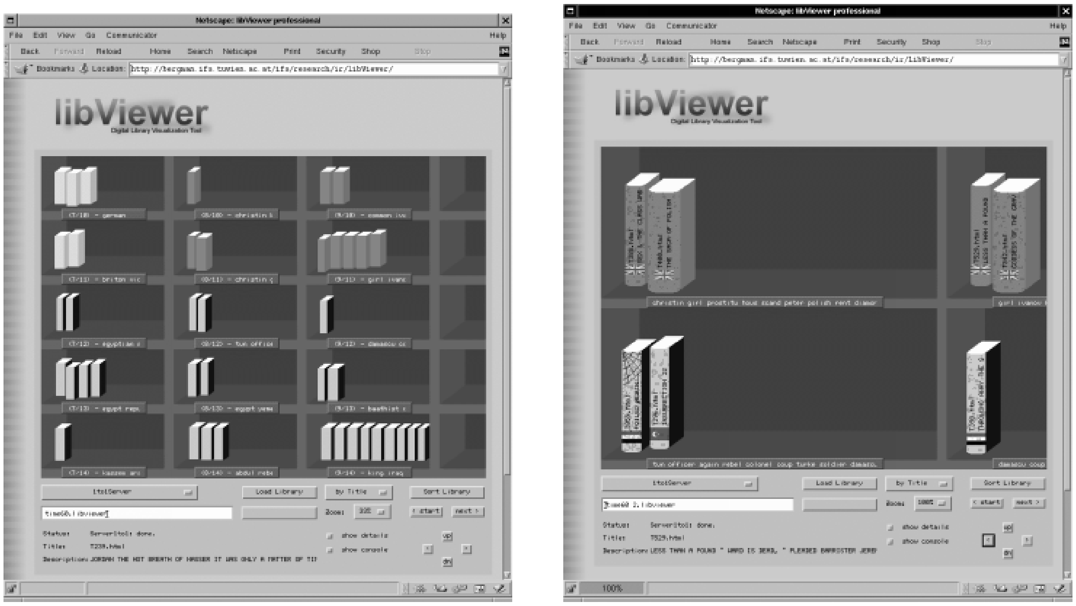
Therefore, we think that the application of the GHSOM in the information retrieval and digital library domain [Rau00] is well-suited because of the inherent hierarchical structure of many document collections as can be seen in the examples described in detail in Chapter 4. The topical hierarchies of a newspaper article collection can be explored in a convenient way due to the relatively small size of the single maps showing only a portion of the data.

One question that could arise is: "What happens, if the document collection is expanded?" In opposition to the example of a static article collection presented in this thesis text collections certainly grow in the course of time. As long as the vocabulary used to describe the current documents, i.e. the index terms, is sufficient for additional documents, the new documents are mapped into the existing hierarchy according to their content. Due to the flexible architecture of the GHSOM continuing the training process of the maps at a later time is possible, so that the maps at the lowest level can grow when too many documents are already represented per unit.

When using the full-text indexing approach described in Section 4.1, the vector space representation has to be renewed if, for example, a completely new topic emerges. Accordingly, the GHSOM has to be re-trained because of the new representation of the documents. Assuming a limited number of words in the world and a constant growth of the document archive, the more documents are present in an archive and the more general the index terms are, the fewer updates of the vector representation have to be performed.

The map metaphor being inherently present in the architecture of the GHSOM might be considered inappropriate in the field of text document classification. A more suitable method to visualize self-organizing maps has been presented in [Rau99a]. The so-called *libViewer* interface uses metaphor graphics to display digital library contents and has been applied to trained SOMs (see Fig. 5.1) as part of the *SOMLib Digital Library System* [Rau99b]. On the left-hand side, in Figure 5.1(a) a bookshelf is depicted representing the spatial organization of documents established during the training process of a *Self-Organizing Map.* Additionally, a set of metaphors (Fig. 5.1(b)) has been implemented to visualize several properties of the documents as we would expect them in a real library, e.g. size (thickness of the spine), frequency of consultation (fingerprints) or the date of last usage (spider webs).

This visualization system can be used to represent the maps of a GHSOM


Netscape: libViewer professional
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bergman.ifs.tuwien.ac.at/ifs/research/ir/libViewer/
libViewer
Digital Library Visualization Tool
(7/14) - kazzem ani
(8/14) - abdul rebe
(9/14) - king iraq
tun officer again rebel colonel coup turks soldier damasc.
christin girl prostitu tous scand peter polish rent diamor
ltolServer
Load Library
by Title
Sort Library
Zoom:
< start
next >
show details
show console
up
dn
100%

Status: Server(tol): done.
Title: T239.html
Description: JORDAN THE HOT BREATH OF NASSER IT WAS ONLY A MATTER OF TI

Status: Server(tol): done.
Title: T529.html
Description: LESS THAN A POUND " WARD IS DEAD, " PLEADED BARRISTER JERD

(a) Bookshelf with various documents

(b) Detailed view of the books

Figure 5.1: **LibViewer Visualization:** The bookshelf metaphor is used to visualize documents organized on a self-organizing map. The location of the books on the shelf correspond to the input vectors mapped onto the according SOM.

hierarchy, whereby the graphical metaphors could be extended to the additional features the GHSOM offers compared to a single, flat SOM. The metaphor of a library building having several floors, multiple rooms on each floor and several bookshelves located in each room could be used to represent the organization of a text collection organized by a GHSOM.

Several other future improvements of the *Growing Hierarchical Self-Organizing Map* are still to be discussed. On issue is a query interface for the GHSOM to allow users to search for and not only browse through interesting documents. One type of query could be a keyword based search to find interesting documents, or the second possibility is to provide a sample text in order to find documents with similar content. In the case of keyword queries it has to be stated that creating the vector which represents the query is a non-trivial task, because due to the sparse allocation of vector elements the adequate weighting of keywords is crucial

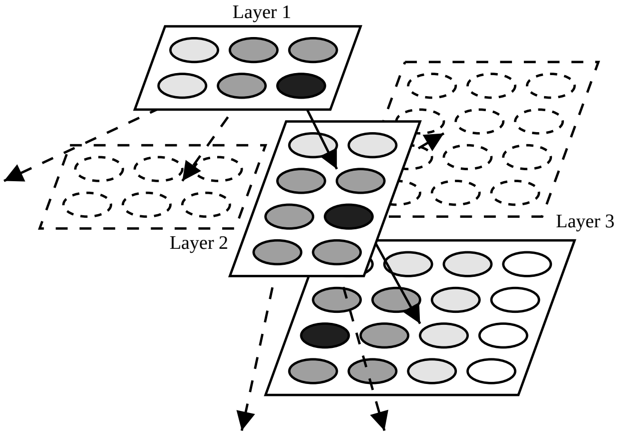


Figure 5.2: **Query Propagation:** A keyword query or sample document, transformed into a vector, can be mapped onto the GHSOM as additional input and traced through the hierarchy to the documents of interest.

for the quality of the query result. If a sample document is provided as query, the *tf* × *idf* weighting scheme can be utilized to create the vector representation.

Consider Figure 5.2 as a part of a trained GHSOM. After mapping the query onto the maps (without any weight vector adaptation), the units could be highlighted according to the similarity of the respective weight vector and the query vector. The best matching unit for a map is depicted as a black circle and units farther away are colored in different shades of gray according to the similarity. A query could then be tracked until the map containing the actual documents is reached.

Another point for further research is to control the orientation of the weight vectors on lower layer maps. In Figure 5.3 a map in one layer and two maps in the next layer, corresponding to two neighboring units in the upper-layer map are depicted.

Consider three topics a, b and c being mapped onto three units in the lower left corner on a map in layer l and two maps exhibiting topics a and b in more detail on two separate maps in layer l+1. Controlling the orientation of the weight vectors means to "ensure" that documents focusing on topic a but including a small amount of information on topic b on one map (on the left-hand side in

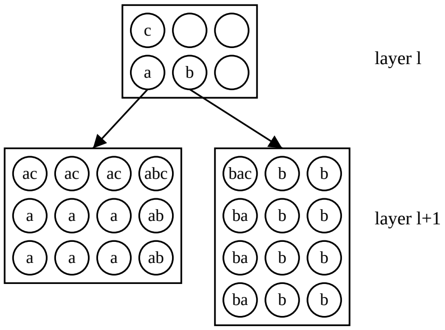


Figure 5.3: **Weight Vector Orientation:** Three different topics (a, b, c) are represented by the respective units in the lower left corner of the map in layer l. The topics are displayed with higher granularity on the maps in layer $l + 1$.

Fig. 5.3) and vice versa on the other map will be mapped spatially close to each other if the maps would be joined hypothetically. This could be achieved by initializing the weight vectors of newly created maps according to the weight vectors of the corresponding unit and its neighbors in the upper layer.

Further ideas for improving the GHSOM are a different criterion for the termination of the growth process of the maps to have an even better representation of the cluster structure of the input data or the execution of the indexing and weighting process for every map in the hierarchy separately, in order to improve the mapping and label quality of the single maps.

To summarize, the *Growing Hierarchical Self-Organizing Map* has shown to be an adequate architecture combining the benefits of other neural network models based on the *Self-Organizing Map* for representing hierarchical structures in data. Its adaptive but nonetheless stable structure built during the training process provides an appropriate interface for interactive data exploration of large amounts of high-dimensional data.

Bibliography

[Bla93] J. Blackmore and R. Miikkulainen. Incremental Grid Growing: Encoding high-dimensional structure into a two-dimensional feature map. In *Proc IEEE Int'l Conf on Neural Networks*, San Francisco, CA, 1993.

[Bau97] H.-U. Bauer and T. Villmann. Growing a hypercubical output space in a self- organizing feature map. *IEEE Transactions on Neural Networks*, 8(2):218–26, 1997.

[Cav94] W. Cavnar. Using an n-gram-based document representation with a vector processing retrieval model. In *Proc. of the Third Text Retrieval Conference (TREC-3)*, pages 269–277, 1994.

[Dee90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Dit00] M. Dittenbach. Results of Data Classification Using the Growing Hierarchical SOM. In A. Rauber and J. Paralic, editors, *Proc. of the Workshop on Data Analysis (WDA 2000)*, Košice, Slovakia, May 26–28 2000.

[Dit00a] M. Dittenbach, D. Merkl, and A. Rauber. The Growing Hierarchical Self-Organizing Map. In S.-I. Amari, C. L. Giles, M. Gori, and V. Puri, editors, *Proc of the International Joint Conference on Neural Networks (IJCNN 2000)*, volume VI, pages 15–19, Como, Italy, July 24–27 2000. IEEE Computer Society.

[Dit00b] M. Dittenbach, D. Merkl, and A. Rauber. Using Growing Hierarchical Self-Organizing Maps for Document Classification. In *Proc of the European Symposium on Artificial Neural Networks (ESANN 2000)*, pages 7–12, Bruges, Belgium, April 26–28 2000. D-Facto Publications.

[Dum94] S. Dumais. Latent semantic indexing (lsi) and TREC-2. In D. K. Harman, editor, *Proc. of The Second Text Retrieval Conference (TREC-2)*, pages 105–115, Gaithersburg, MD, March 1994. NIST. Special publication 500-215.

[Fri94] B. Fritzke. Growing cell structures – a self-organizing network for unsupervised and supervised learning. *Neural Networks*, 7(9):1441–1460, 1994.

[Fri95] B. Fritzke. Growing grid – a self-organizing network with constant neighborhood range and adaption strength. *Neural Processing Letters*, 2(5):9–13, 1995.

[Fri96] B. Fritzke. Growing self-organizing networks — Why? In *Proc Europ Symp on Artificial Neural Networks (ESANN 96)*, Bruges, Belgium, 1996.

[Him00] J. Himberg. A SOM based cluster visualization and its application for false coloring. In S.-I. Amari, C. L. Giles, M. Gori, and V. Puri, editors, *Proc. of the International Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, July 24–27 2000. IEEE Computer Society Press.

[Hon97] T. Honelka, S. Kaski, K. Lagus, and T. Kohonen. WEBSOM – Self-Organizing Maps of Document Collections. In *Workshop on Self-Organizing Maps (WSOM 97)*, Espoo, Finland, June 4–6 1997.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 27:417–441, 1933.

[Jai99] A.K. Jain, M.N. Murty, and P.J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[Kas98] S. Kaski. Dimensionality reduction by random mapping. In *Proc. of the Int'l. Joint Conference on Neural Networks (IJCNN 98)*, pages 413–418, Piscataway, NJ, 1998. IEEE Press.

[Kas96] S. Kaski and T. Kohonen. Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World. In *Proc. of 3rd Int'l. Conference on Neural Networks in the Capital Markets*, pages 498–507, London, England, October 11–13 1996. World Scientific, Singapore.

[Kas98a] S. Kaski, J. Kangas, and T. Kohonen. Bibliography of Self-Organizing Map (SOM) Papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176, 1998.

[Köh96] M. Köhle and D. Merkl. Identification of gait patterns with self-organizing maps based on ground reaction force. In *Proc. of the European Symposium on Artificial Neural Networks (ESANN 96)*, Bruges, Belgium, 1996.

[Koh82] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.

[Koh89] T. Kohonen. *Self-Organization and Associative Memory.* Springer-Verlag, Berlin, 3rd edition, 1989.

[Koh95] T. Kohonen. *Self-organizing maps.* Springer-Verlag, Berlin, 1995.

[Kas99] S. Kaski, J. Venna, and T. Kohonen. Coloring that reveals high-dimensional structures in data. In *Proc. of the 6th International Conference on Neural Information Processing (ICONIP 99)*, volume II, pages 729–734, Piscataway, NJ, 1999. IEEE Service Center.

[Lag96] 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.

[Laa99] J. Laaksonen, M. Koskela, and E. Oja. PicSOM – A Framework for Content-Based Image Database Retrieval using Self-organizing Maps. In *Proc. of the 11th Scandinavian Conference on Image Analysis (SCIA 99)*, Kangerlussuaq, Greenland, June 7–11 1999.

[Lin91] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proc. of the Int'l ACM SIGIR Conf on R&D in Information Retrieval (SIGIR 91)*, Chicago, IL, 1991.

[Mer97a] 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.

[Mer97b] D. Merkl. Lessons learned in text document classification. In *Workshop on Self-Organizing Maps (WSOM 97)*, pages 316–321, Espoo, Finland, June, 4–6 1997.

[Mii90] R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2, 1990.

[Mer97] D. Merkl and A. Rauber. Alternative ways for cluster visualization in self-organizing maps. In *Workshop on Self-Organizing Maps (WSOM 97)*, Espoo, Finland, June, 4–6 1997.

[Mer00] D. Merkl and A. Rauber. Uncovering the Hierarchical Structure of Text Archives by Using an Unsupervised Neural Network with Adaptive Architecture. In *Proc. of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2000)*, pages 384–395, Kyoto, Japan, April 18–20 2000. Springer Lecture Notes in Artificial Intelligence (LNCS-LNAI 1805).

[Rau99] A. Rauber. LabelSOM: On the Labeling of Self-Organizing Maps. In *Proc of the International Joint Conference on Neural Networks (IJCNN 99)*, 1999.

[Rau99a] A. Rauber and H. Bina. A Metaphor Graphics Based Representation of Digital Libraries on the World Wide Web: Using the libViewer to Make Metadata Visible. In *Workshop Proc. of the 10th Intl. Conf. on Database and Expert Systems Applications, Workshop on Web-Based Information Visualization (WebVis 99)*, Florence, Italy, August 30 – September 3 1999. IEEE Press.

[Rau00] A. Rauber, M. Dittenbach, and D. Merkl. Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural Network Approach to Bookshelf Creation and Arrangement. In J. Borbinha and T. Baker, editors, *Proc. of the Fourth Europ. Conf. on Research and Advanced Technology for Digital Libraries(ECDL 2000)*, number 1923 in LNCS, pages 348–351, Lisbon, Portugal, September 18–20 2000. Springer-Verlag Berlin.

[Rau98] A. Rauber and D. Merkl. Finding structure in text archives. In *Proc. European Symp. on Artificial Neural Networks (ESANN 98)*, Bruges, Belgium, 1998.

[Rau99b] A. Rauber and D. Merkl. The SOMlib Digital Library System. In *Proc. of the Third Europ. Conf. on Research and Advanced Technology for Digital Libraries (ECDL 99)*, number 1696 in LNCS, Paris, France, September 22–24 1999. Springer-Verlag Berlin.

[Sal89] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison Wesley, Reading, MA, 1989.

[Sch99] E. Schweighofer and D. Merkl. A learning technique for legal document analysis. In *Proc. of the International Conference on Artificial Intelligence and Law*, Oslo, Norway, 1999.

[Sch00] E. Schweighofer, A. Rauber, and D. Merkl. Some remarks on vector representation of legal documents. In R.R. Wagner, A M. Tjoa and A. Al-Zobaidie, editors, *DEXA Workshop Proceedings of the Work-*

shop on Legal Information Systems (LISA 2000), pages 1087 – 1091, Greenwich, UK, Sept. 4. – 8. 2000. IEEE Computer Society Press.

[Ult93] A. Ultsch. Self-organizing neural networks for visualization and classification. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 307–313. Springer-Verlag, Berlin, 1993.