The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data

Andreas Rauber, *Member, IEEE*, Dieter Merkl, *Member, IEEE*, and Michael Dittenbach

***Abstract*—The self-organizing map (SOM) is a very popular unsupervised neural-network model for the analysis of high-dimensional input data as in data mining applications. However, at least two limitations have to be noted, which are related to the static architecture of this model as well as to the limited capabilities for the representation of hierarchical relations of the data. With our novel growing hierarchical SOM (GHSOM) presented in this paper, we address both limitations. The GHSOM is an artificial neural-network model with hierarchical architecture composed of independent growing SOMs. The motivation was to provide a model that adapts its architecture during its unsupervised training process according to the particular requirements of the input data. Furthermore, by providing a global orientation of the independently growing maps in the individual layers of the hierarchy, navigation across branches is facilitated. The benefits of this novel neural network are a problem-dependent architecture and the intuitive representation of hierarchical relations in the data. This is especially appealing in explorative data mining applications, allowing the inherent structure of the data to unfold in a highly intuitive fashion.**

***Index Terms*—Data mining, exploratory data analysis, hierarchical clustering, pattern recognition, self-organizing map (SOM).**

I. Introduction

DATA MINING, or more generally, pattern recognition and knowledge acquisition, heavily depend on suitable unsupervised learning methods. The purpose 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, which are 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 to a pattern belonging to a different cluster [1]. 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 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) is being widely used as a tool for mapping high-dimensional data into a two-dimensional (2-D) representation space [2]. This mapping retains the relationship between input data as faithfully as possible, thus describing a topology-preserving representation of input similarities in terms of distances in the output space. It is then possible to visually identify clusters on the map. The main advantage of such a mapping is the ease by which a user gains an idea regarding the structure of the data by analyzing the map.

However, some difficulties in SOM utilization remained largely untouched, even though a large number of research papers on applications of the SOM were presented over the years [3]. First, the SOM uses a fixed network architecture in terms of number and arrangement of neural processing elements, which has to be defined prior to training. Obviously, in case of largely unknown input data characteristics, it remains far from trivial to determine the network architecture that provides satisfying results. Thus, it certainly is worth considering neural-network models that determine the number and arrangement of units during their unsupervised training process. We refer to [4]–[6] for recently proposed models that are based on the SOM, yet allow for adaptation of the network architecture during training.

Second, hierarchical relations between the input data are not mirrored in a straightforward fashion. Such relations are, rather, shown within the same representation space and, thus, are hard to identify. Hierarchical relations, however, may be observed in a wide spectrum of application domains. Thus, their proper identification remains a highly important data mining task that cannot be addressed conveniently within the framework of the SOM.

To resolve both limitations of the SOM in a uniform fashion we propose a novel artificial neural-network architecture in this paper, namely the growing hierarchical SOM (GHSOM). This model has a hierarchical architecture, where SOM-like neural networks with adaptive architecture build the various layers of the hierarchy. The size of these SOM-like neural networks as well as the depth of the hierarchy of the GHSOM is determined during its unsupervised training process according to the structure of the data.

The hierarchical structuring imposed on the data results in a separation of clusters mapped onto different branches. While this, in principle, is a desirable characteristic helping to understand the cluster structure of the data, it may lead to misinterpretations when large clusters are mapped onto and expanded from two neighboring, yet different units. Similar input data might, thus, be rather arbitrarily separated into different branches of the hierarchy. By choosing the initial orientation of deeper layers according to their respective higher layer maps, we can maintain the already learned similarities between input data during the creation of the hierarchical structure of the GHSOM. As a consequence, the negative effects of generating strictly disjoint clusters are eliminated

Manuscript received December 1, 2000.

The authors are with the Department of Software Technology and Interactive Systems, Vienna University of Technology, A-1040 Vienna, Austria (e-mail: andi@ifs.tuwien.ac.at; dieter@ifs.tuwien.ac.at; mbach@ifs.tuwien.ac.at).

Digital Object Identifier 10.1109/TNN.2002.804221

1045-9227/02$17.00 © 2002 IEEE

because neighboring maps in deeper layers of the hierarchy show similar characteristics at their respective borders.

We show the usefulness of the GHSOM 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 *TIME* Magazine, covering a variety of topics ranging from international politics to social gossip. 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 paper is structured as follows: Section II provides an introduction to the architecture and training process of the SOM, followed by a review of related architectures in Section III. The GHSOM architecture is introduced and presented in detail in Section IV. Two different data sets are used to demonstrate the characteristics and capabilities of the GHSOM model in Section V, starting with the smaller *TIME* Magazine collection in Section V-A, followed by the more extensive collection of articles from the newspaper *Der Standard* in Section V-B. Some remarks conclude the paper in Section VI.

II. SOM

The SOM, as proposed in [2] and described thoroughly in [7]–[9], is one of the most distinguished artificial neural-network models adhering to the unsupervised learning paradigm. The SOM is a general unsupervised tool for the ordering of high-dimensional data in such a way that similar items are grouped spatially close to one another.

The range of applications where the SOM has been utilized successfully is impressive; see [3] for a fairly recent bibliography. The model also has a strong tradition in the text mining area where a number of research groups described work based on the SOM [10]–[14].

The SOM consists of a number of neural processing elements, i.e., units that are arranged according to some topology, the most common choice of which is marked by a 2-D rectangular or hexagonal grid. Each of the units i is further assigned a model vector m_i, $m_i \in \Re^n$. It is important to note that these model vectors have the same dimensionality as the input patterns.

The training process of SOMs may be described in terms of input pattern presentation and model vector adaptation. Each training iteration t starts with the random selection of one input pattern x, $x \in \Re^n$. This pattern is presented to the SOM and each unit determines its activation. Usually, the Euclidean distance between input pattern and model vector is used to calculate a unit's activation. In this case, the unit having the model vector with the smallest Euclidean distance to the input pattern is referred to as the *winner*. We will use the index c for denoting the *winner* [cf. (1)]

$$c(t) = \arg\min_i \{||x(t) - m_i(t)||\}. \qquad (1)$$

Finally, the model vector of the *winner* as well as model vectors of units in the vicinity of the *winner* are adapted. This adaptation is implemented as a gradual reduction of the difference between corresponding components of the input pattern and the model vector, as shown in (2). Note that we make use of discrete-time notation with t denoting the current training iteration

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

Geometrically speaking, the model vectors of the adapted units are moved a bit toward the input pattern. The amount of model vector movement is guided by a learning rate α, decreasing in time. The number of units that are affected by adaptation as well as the strength of adaptation depending on a unit's distance from the *winner* is determined by a neighborhood function h_{ci}. This number of units also decreases in time such that toward the end of the training process only the *winner* is adapted. Typically, the neighborhood function is a unimodal function which is symmetric around the location of the winner and monotonically decreasing with increasing distance from the winner. Often, a Gaussian is used as a neighborhood function as given in

$$h_{ci}(t) = \exp\left(-\frac{||r_c - r_i||^2}{2 \cdot \delta(t)^2}\right). \qquad (3)$$

In this expression, $||r_c - r_i||^2$ denotes the distance between units c and i within the output space, with r_i representing the 2-D location vector of unit i within the grid. The time-varying parameter δ guides the reduction of the neighborhood kernel during training. It is common practice that this neighborhood kernel is selected large enough to cover a wide area of the output space in the beginning of learning. The spatial width of the kernel is reduced gradually during training such that toward the end of the process just the *winner* is adapted.

III. Related Architectures

During recent years, a number of modifications have been suggested to enhance the usefulness of the SOM for data mining applications. In particular, the identification of inter- and intra-cluster similarity has been addressed. Approaches such as the U-Matrix [15], adaptive coordinates, and cluster connections [16] represent techniques that put emphasis on detection and visualization of cluster structures in SOMs. These techniques analyze the distances between neighboring units or mirror the effect of the model vector adaptation in the 2-D output space. Similar cluster information can be obtained with our LabelSOM method described in [17] and [18]. Using the LabelSOM method, the characteristics of the various units are described in terms of shared features among the input patterns mapped onto a particular unit. Grouping units that have the same characteristics allows to identify clusters within the output space of the SOM. Other modifications of the architecture address the problems arising from the mapping onto a lattice structure, providing a smooth manifold as output space, such as the self-organizing field [19], or the generative topographic mapping (GTM) [20]. However, none of the methods identified above facilitates the detection of hierarchical structure inherent in the data or adapt the size of the network.

The hierarchical feature map [21] tries to uncover the hierarchical structure of data by modifying the SOM architecture.

Instead of training a flat SOM, a balanced hierarchical structure of SOMs is trained. Data mapped onto one single unit is represented at a further level of detail in the lower level map assigned to this unit. However, this model merely represents the data in a hierarchical way, rather than really reflecting the hierarchical structure of the data. This is because the architecture of the network has to be defined in advance, i.e., the number of layers and the sizes of the maps at each layer are fixed prior to network training. This leads to the definition of a balanced tree which is used to represent the data. Desirable is, however, a network architecture defined by the peculiarities of the input data.

In [22] and [23], a tree-structured SOM, a hierarchical modification of the SOM, is introduced. The focus in this model, however, is rather on computational speedup during *winner* selection by using a tree-based organization of the units. This model does not provide a hierarchical decomposition of the input space; the input patterns are, again, organized in one single flat SOM.

The shortcoming of having to define the size of the SOM in advance has been addressed by a number of different models. Consider, for example, the incremental grid growing [4], growing grid [5], growing SOM [6], and the hypercubical SOM [24]. Incremental grid growing allows the addition of new units at the boundary of the map. Furthermore, connections between units of the map may be established and removed according to some threshold settings based on the similarity of their respective model vectors. This may result in several separated irregularly shaped map structures. The growing SOM, quite similar in spirit, uses a spread factor to control the growth process of the map. Using manual intervention, the supervisor can decide to train separate SOMs for specific units in order to obtain a more detailed representation. This obviously results in manually created hierarchies, an approach that is basically possible with each variant of the SOM.

The growing grid, on the other hand, adds rows and columns of units during the training process, starting with a SOM of initially 2 × 2 units. The decision where to insert new units is governed by the computation of some measure for each unit, e.g., the winner counter. As an extension, the hypercubical SOM allows for a growth process of the SOM into more than two dimensions, thus providing improved data representation while forsaking visual interpretability.

It can be stated, however, that the main focus of each of these adaptive variants of the SOM lies with an equal distribution of the input patterns across the map by adding new units in the neighborhood of units that represent an unproportionally high number of input data. Thus, they do not primarily reflect the concept of representation at a certain level of detail, which is expressed in terms of the overall quantization error rather than in the number of input data mapped onto specific areas. Moreover, neither of these adaptive models takes the inherently hierarchical structure of data into account.

IV. GHSOM

A. Principles

While the SOM has proven to be a very suitable tool for detecting structure in high-dimensional data and organizing it accordingly on a 2-D output space, some shortcomings have to be mentioned. These include its inability to capture the inherent hierarchical structure of data. Furthermore, the size of the map has to be determined in advance when proper insight into the characteristics of a data distribution might not be available. These drawbacks have been addressed separately in several modified architectures of the SOM as outlined in Section III. However, none of these approaches provides an architecture which fully adapts to the characteristics of the input data. To overcome the limitations of both fix-sized and nonhierarchically adaptive architectures, we developed the GHSOM, which dynamically fits its multilayered architecture according to the structure of the data [25].

The GHSOM has a hierarchical structure of multiple layers, where each layer consists of several independent growing SOMs. Starting from a top-level map, each map, similar to the growing grid model, grows in size to represent a collection of data at a specific level of detail. After a certain improvement regarding the granularity of data representation is reached, the units are analyzed to see whether they represent the data at a specific minimum level of granularity. Those units that represent too diverse input data are expanded to form a new small growing SOM at a subsequent layer, where the respective data shall be represented in more detail. These new maps again grow in size until a specified improvement of the quality of data representation is reached. Units representing an already rather homogeneous set of data, on the other hand, will not require any further expansion into subsequent layers. The resulting GHSOM, thus, is fully adaptive to reflect, by its very architecture, the hierarchical structure inherent in the data, allocating more space for the representation of inhomogeneous areas in the input space.

A graphical representation of a GHSOM is given in Fig. 1. The map in layer 1 consists of 3 × 2 units and provides a rather rough organization of the main clusters in the input 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 from one of the second-layer maps have further been expanded into third-layer maps to provide sufficiently granular input data representation. It has to be noted that the maps have different sizes according to the structure of the data, which relieves us from the burden of predefining the structure of the architecture. The layer 0 serves as a representation of the complete data set and is necessary for the control of the growth process.

B. Training Algorithm

1) *Initial Setup and Global Network Control:*

The principle of the GHSOM architecture is its adaptation to the training data. The quality of this adaptation is measured in terms of the deviation between a unit's model vector and the input vectors represented by this particular unit. Basically, two different strategies can be used for the control of the growth process, using either the mean quantization error (mqe) of a unit (which is commonly used as a quality measure for data representation with SOMs), or the absolute value, i.e., the quantization error (qe) of a unit.

More formally, the mqe of a unit i is calculated according to (4) as the mean Euclidean distance between its model vector m_i

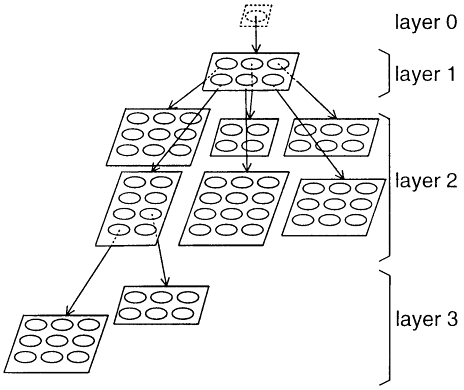


Fig. 1. GHSOM: The GHSOM evolves to a structure of SOMs reflecting the hierarchical structure of the input data.

and the $n_{\mathcal{C}}$ input vectors x_j that are elements of the set of input vectors $\mathcal{C}_i$ mapped onto this unit i

$$\text{mqe}_i = \frac{1}{n_{\mathcal{C}}} \cdot \sum_{x_j \in \mathcal{C}_i} \|m_i - x_j\|, \qquad n_{\mathcal{C}} = |\mathcal{C}_i|,\, \mathcal{C}_i \neq \emptyset. \quad (4)$$

The starting point for the GHSOM training process is the calculation of an mqe_0 of the unit forming the layer 0 map as provided in (5). With $n_{\mathcal{I}}$ we refer to the number of all input vectors x of the input data set $\mathcal{I}$ and m_0 denotes the mean of the input data

$$\text{mqe}_0 = \frac{1}{n_{\mathcal{I}}} \cdot \sum_{x_i \in \mathcal{I}} \|m_0 - x_i\|, \qquad n_{\mathcal{I}} = |\mathcal{I}|. \quad (5)$$

The mqe measures the dissimilarity of all input data mapped onto a particular unit and will be used to control the growth process of the neural network. Specifically, the minimum quality of data representation of each unit will be specified as a fraction, indicated by a parameter τ_2, of mqe_0.

In other words, all units must represent their respective subsets of data at an mqe smaller than a fraction τ_2 of mqe_0, i.e., satisfy the global termination criterion specified in (6)

$$\text{mqe}_i < \tau_2 \cdot \text{mqe}_0. \quad (6)$$

For all units not satisfying this condition, a more detailed data representation is required, leading to the addition of further units to provide more map space for data representation.

Alternatively, the quantization error (qe) of the unit as given in (7), denoted by qe, may be used instead of the mqe, resulting in a global termination criterion as follows:

$$\text{qe}_i = \sum_{x_j \in \mathcal{C}_i} \|m_i - x_j\| \quad (7)$$

$$\text{qe}_i < \tau_2 \cdot \text{qe}_0. \quad (8)$$

While using the mqe of the data distribution as a global quality measure for the GHSOM training process may be more intuitive, using the qe follows more closely the principle characteristic of SOMs of providing more map space for more densely populated regions of the input space, also referred to as *magnification factor*. Thus, using (8) rather than (6) as a global stopping criterion, we produce maps that more intuitively reflect the characteristics of data distributions by capturing finer differences in more densely populated clusters. This effect is specifically important when the resulting maps shall be used as explorative interfaces to data sets, as it is frequently the case for SOM-like architectures. We will thus, for the remainder of this paper, use the qe as a basis for GHSOM training.

Following the decision whether to use mqe or qe for the global control of the training process, the GHSOM architecture is initialized by creating a new growing SOM beneath the layer 0 map. The initial size of this first-layer map is set to 2 × 2 units, with its model vectors being initialized to random values.

2) Training and Growth Process of a Growing SOM: A newly created map is trained according to the standard SOM training procedure as described in Section II. After a fixed number λ of training iterations, the qes of all units as provided in (7) are analyzed. A high qe indicates that an inhomogeneous part of the input space containing dissimilar data, or at least a rather large set of input data from a more homogenous part of the input space is represented by this unit. Therefore, new units are needed to provide more space for appropriate data representation. Thus, the unit with the highest qe is selected and denoted as the *error unit*. We will refer to the error unit as e. Next, the most dissimilar neighboring unit d in terms of input space distance is selected. This is done by comparing the model vectors of all neighboring units with the model vector of the error unit e. A new row or column of units is inserted between e and its most dissimilar neighbor d. The model vectors of the new units are initialized as the average of their corresponding neighbors.

Fig. 2 shows a graphical representation of the insertion process of our realization of a growing SOM, with the newly inserted units being depicted as shaded circles. The arrows point to the respective neighboring units used for model vector initialization.

More formally, the growth process of a growing SOM can be described as follows. Let $\mathcal{C}_i$ be the subset of vectors x_j of the input data that is mapped onto unit i, i.e., $\mathcal{C}_i \subseteq \mathcal{I}$; and m_i the model vector of unit i. Then, the error unit e is determined as the unit with the mqe as follows:

$$e = \arg\max_i \left(\sum_{x_j \in \mathcal{C}_i} \|m_i - x_j\| \right), \qquad n_{\mathcal{C}} = |\mathcal{C}_i|,\, \mathcal{C}_i \neq \emptyset. \quad (9)$$

Please note, that the mqe may be used instead of the qe, resulting in a GHSOM architecture focusing on overall homogeneity of data representation, rather than capturing higher degrees of detail for more densely populated areas of the data space.

Following the selection of e, its most dissimilar neighbor d is determined as follows, where $\mathcal{N}_e$ is the set of neighboring units of e

$$d = \arg\max_i (\|m_e - m_i\|), \qquad m_i \in \mathcal{N}_e. \quad (10)$$

A row or column of units is inserted between d and e. To obtain a smooth positioning of the newly added units in the input space, their model vectors are initialized as the means of their respective neighbors. After insertion, the learning rate and neigh-

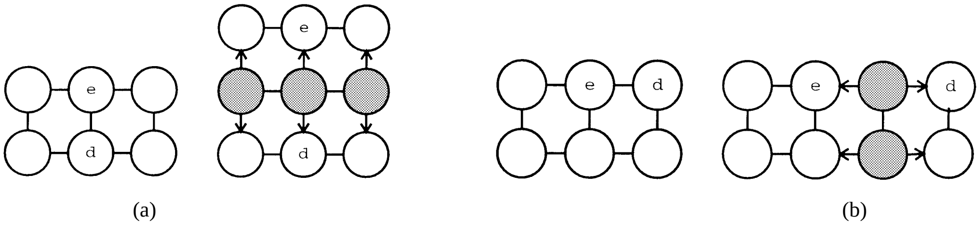


Fig. 2. Insertion of units: (a) A row or (b) a column of units (shaded gray) is inserted in between *error unit* e and the neighboring unit d with the largest distance between its model vector and the model vector of e in the Euclidean space.

borhood range are reset to their original values, and training continues in a SOM-like fashion for the next λ iterations.

This training process of single growing SOM is highly similar to the growing grid model [5]. The difference, so far, is that we use a decreasing learning rate and a decreasing neighborhood range instead of fixed values.

3) Termination of Growth Process: As more units are added to the growing SOM their qes decrease as each of the units represents a smaller, more concise subset of the input space. Basically, the training process could continue until all units satisfy the global stopping criterion, i.e., they represent their respective subset of the data at a granularity lower than a certain fraction of the initial standard deviation of the data, resulting in a large SOM representing all data at the required granularity in one layer.

Yet, in order to reveal the hierarchical structure present in the data, each map shall only explain a portion of data similarity. Thus, the growth process continues only until the map's mqe, referred to as MQE in capital letters, reaches a certain fraction τ_1 of the qe_u of the corresponding unit u in the upper layer (i.e., the unit constituting the layer 0 map for the first-layer map). The MQE of a map is computed as the mean of all units' quantization errors qe_i [cf. (7)] of the subset $\mathcal{U}$ of the maps' units onto which data is mapped

$$\mathrm{MQE}_m = \frac{1}{n_{\mathcal{U}}} \cdot \sum_{i \in \mathcal{U}} \mathrm{qe}_i, \qquad n_{\mathcal{U}} = |\mathcal{U}|. \tag{11}$$

In general terms, the stopping criterion for the growth of a single map m is defined as follows:

$$\mathrm{MQE}_m < \tau_1 \cdot \mathrm{qe}_u \tag{12}$$

where qe_u is the qe of the corresponding unit u in the upper layer. Obviously, the smaller the parameter τ_1 is chosen the larger the resulting map will be, explaining its data at a higher granularity. For larger τ_1, more detailed data representation will be delegated to additional maps further down the hierarchy. Thus, the parameter τ_1 serves as the control parameter for the depth/shallowness of the resulting hierarchical GHSOM architecture.

In the case of the first-layer map, the stopping criterion for the training process is $\mathrm{MQE}_1 < \tau_1 \cdot \mathrm{qe}_0$.

4) Hierarchical Growth With Global Map Orientation: When the training of a map is finished according to the criterion specified in (12), every unit has to be checked for fulfillment of the global stopping criterion given in (8). Units representing a set of too diverse input vectors, are expanded to form a new map at a subsequent layer of the hierarchy. Units satisfying the global stopping criterion require no further expansion.

Similar to the procedure chosen for the creation of the layer 1 map, originating from the single-unit map at layer 0, a new map of initially 2 × 2 units is created. While, again, random initialization can be chosen for model vector orientation, this usually will distort the global topology of neighboring maps. This is because the orientation of data on this map, i.e., which subclusters are located on which area of the map, is determined by the self-organizing process following the random orientation of the map in data space. Thus, navigation across maps within the same layer of the hierarchy would be prevented, leading to serious disadvantages of the organization of disjoint clusters in a hierarchical manner, especially when larger clusters should be split. To provide a global orientation of the individual maps in the various layers of the hierarchy, their orientation must conform to the orientation of the data distribution on their parent's map. This can be achieved by creating a coherent initialization of the units of a newly created map [26].

Let unit p be expanded to form a new 2 × 2 map in the subsequent layer of the hierarchy. This map's four model vectors $s1$ to $s4$ are initialized to mirror the orientation of neighboring units of its parent p. Fig. 3 provides an illustration of the initialization of new maps and the influence of the parent unit's neighbors. Geometrically speaking, the model vectors of the four corner units are moved in data space toward the directions of their respective parent's neighbors by a certain fraction. This initial orientation of the map is preserved during the training process. While new units may be inserted in between, the four corner units will still be most similar to the respective corner units of the maps in neighboring branches. The exact amount by which these corner units are moved in the respective directions does not influence the characteristic of the topology-preserving initialization. Thus, we can choose to set them to the mean of the parent and its neighbors in the respective directions, e.g., setting $e1 = (e + a + b + d)/4$. In the simplest case, the neighbors' model vectors may be used directly as initial corner positions of the new maps in data space.

The input vectors to train the newly added map are the ones mapped onto the unit which has just been expanded, i.e., the subset of the data space mapped onto its parent. This map will again continue to grow following the procedures detailed in Section IV-B2. The whole process is repeated for the subsequent layers until the global stopping criterion given in (8) is met by all leaf units.

5) Analysis of GHSOM Characteristics: The SOM offers itself for the analysis of large high-dimensional data collections, allowing many shortcuts to accelerate the training process, see

[10] for a concise treatment. Additionally, due to its hierarchical structuring and partitioning of the input data, the GHSOM provides improved scalability. At the transition from one layer to the next, the number of input vectors used for training a particular map decreases to the subset of vectors mapped onto the respective upper layer unit.

Furthermore, each map in the hierarchy explains a particular set of characteristics of its input data. Subsequently, some input vector components, i.e., features of the data set, can be expected to be almost identical for all input vectors mapped onto a specific unit. These features can be ignored at the transition to a subsequent layer of the hierarchy, allowing to shorten the input vectors. Especially for very high-dimensional and sparse data sets this effect allows for a significant reduction of computational effort.

The training and growth process of the GHSOM is entirely data driven, requiring no prior knowledge or estimates for parameter specification. The hierarchical structure of data can be represented in different forms, favoring either: 1) lower hierarchies with rather detailed refinements presented at each subsequent layer or 2) deeper hierarchies, which provide a stricter separation of the various subclusters by assigning separate maps. Parameter τ_1 is used to control this tradeoff between shallow or deep hierarchies.

In the first case, we will prefer larger maps in each layer, which explain larger portions of the data in their flat representation, yet providing less hierarchical structuring. As an extreme example, we might consider the growing grid, which grows in size explaining the complete structure of the data in one single flat map. It ignores all hierarchical information and tries, at best, to preserve it in the mapping of various clusters on the flat structure. On the other hand, we might consider setting τ_1 rather large, which requires only limited growth of individual maps, resulting in a deeper hierarchical structure of small maps focusing on the 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.

In principle, the choice of τ_1 may seem crucial, as it might result in a rather arbitrary separation of, a larger cluster with homogeneous data distribution into two or more subclusters in different branches. However, due to the global orientation provided by the initialization of the model vectors of new maps, navigation across maps in the same layer of the hierarchy is facilitated, offsetting the damage of cluster separation. Furthermore, by analyzing the distances in input space of model vectors on neighboring map boundaries, the similarity of neighboring maps can be detected and indicated.

For exploratory data analysis, a homogeneous distribution of data samples across the map space is desired, allowing to capture finer differences between clusters in more densely populated areas of the data space. Thus, using the qe of a unit, rather than its mqe, has shown to produce more favorable results. The global stopping criterion, be it an absolute value or a fraction specified by parameter τ_2, directly influences the overall size of the resulting GHSOM, i.e., the number of units available for data space representation.

It should further be noted that the training process usually does not lead to a balanced hierarchy in terms of all branches having the same depth. This is one of the main advantages of the GHSOM over the hierarchical feature map [21], 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 create deeper branches than others.

V. Experiments

For the following experiments, we use an information retrieval application as a testbed for the GHSOM. In a nutshell, topical clusters shall be detected in a collection of free-form documents with documents covering similar topics to be grouped together. This application domain represents an ideal and challenging scenario for clustering algorithms, as typically very high-dimensional feature spaces are involved. Furthermore, the data can be considered highly noisy as a result from the indexing process that is used to approximately capture the content of a particular document.

Two different experimental settings are presented, focusing on different characteristics of the GHSOM. In Section V-A the classic *TIME* Magazine collection is used to compare the characteristics and features of the hierarchical structuring of a data collection with respect to an SOM. In Section V-B, we analyze the characteristics of shallow and deep hierarchies depending on various settings of parameter τ_1, and present the benefits of the preservation of the global orientation in the GHSOM hierarchy. For these experiments, we use a larger collection of news articles from the Austrian newspaper *Der Standard*. All experiments presented in this paper, as well as an implementation of the presented GHSOM architecture are available at the SOMLib project homepage[1] for interactive evaluation.

A. Experiment 1: TIME Magazine

1) Data Representation: In the first step, the documents have to be mapped into some representation language in order to enable further analysis. This process is termed indexing in the information retrieval literature. A number of different strategies have been suggested over the years of information retrieval research. Still one of the most common representation techniques is single term full-text indexing, where the text of the documents is accessed and the various words forming the document are extracted. These words can be reduced to their word stem yielding the terms used to represent the documents. The resulting set of terms is further cleared from stop-words, i.e., words that appear either too often or too rarely within the document collection and, thus, have only little influence on the discrimination between different documents and would just unnecessarily increase the computational load during classification.

In the vector-space model of information retrieval the documents contained in a collection are represented by means of feature vectors x of the form $x = [\xi_1, \xi_2, \ldots, \xi_n]^T$. In such a representation, the i, $1 \le i \le n$, correspond to the index terms extracted from the documents as described above. The specific

[1] [Online] Available: http://www.ifs.tuwien.ac.at/~andi/somlib.

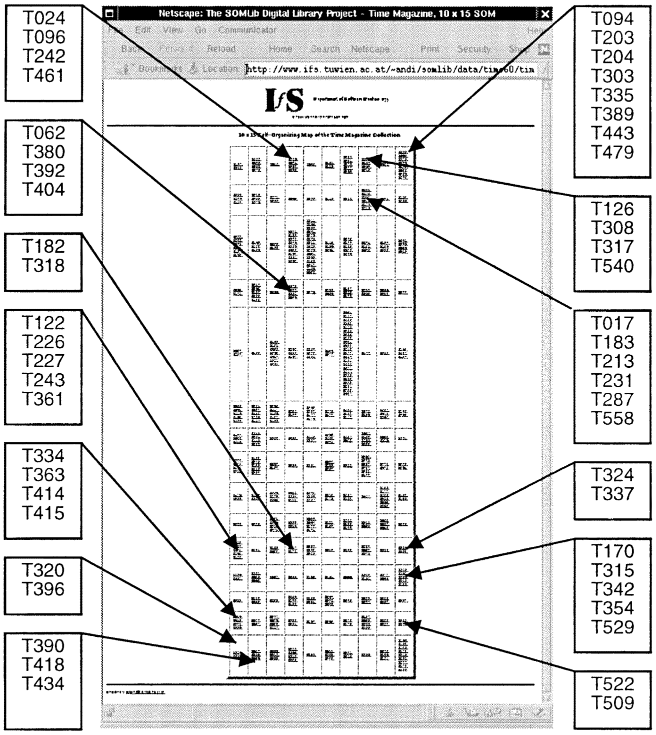

Netscape: The SOMLib Digital Library Project - Time Magazine, 10 x 15 SOM
http://www.ifs.tuwien.ac.at/~andi/somlib/data/time60/tim
IfS
T024 T096 T242 T461
T062 T380 T392 T404
T182 T318
T122 T226 T227 T243 T361
T334 T363 T414 T415
T320 T396
T390 T418 T434
T094 T203 T204 T303 T335 T389 T443 T479
T126 T308 T317 T540
T017 T183 T213 T231 T287 T558
T324 T337
T170 T315 T342 T354 T529
T522 T509

Fig. 4. *TIME* Magazine SOM: 10 × 15 map of the *TIME* Magazine collection.

value of ξ_i corresponds to the importance of index term i in describing the content of the particular document at hand. One might find a lot of strategies to prescribe the importance of an index term for a particular document [27]. Without loss of generality, we may assume that this importance is represented as a scalar in the range of [0, 1], where 0 means that this particular index term is absolutely unimportant to describe the document. Any deviation from 0 toward 1 is proportional to the increased importance of the index term at hand. In such a vector-space model, the similarity between two documents corresponds to the similarity of their vector representations [28].

The *TIME* Magazine article collection consists of 420 articles from the *TIME* magazine of the 1960s, 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 are roughly stemmed and weighted according to a $tf \times idf$, i.e., term frequency times inverse document frequency, weighting scheme [29]. This weighting scheme assigns high values to terms that appear frequently within one document, yet rarely within the overall document collection, i.e., to terms that are considered important in describing the contents of a document. Following the feature extraction process, we end up with 420 vectors describing the documents in the 5923-dimensional feature space. These vectors are further used for neural-network training.

2) SOM of the TIME Magazine Collection: Fig. 4 shows an SOM trained with the *TIME* Magazine data set. It consists of 10 × 15 units represented as table cells with a number of articles being mapped onto each individual unit. The articles mapped onto the same or neighboring units are considered to be similar to each other in terms of the topic they deal with. Due to space considerations, we cannot present all the articles in the collection. Thus, we selected a number of units for detailed discussion.

We find that the SOM has succeeded in creating a topology preserving representation of the topical clusters of articles. For example, in the lower left corner, we find a group of units representing articles on the conflict in Vietnam. To name a few, we find articles *T320, T369* on unit (1/15),[2] or *T390, T418, T434* on the neighboring unit (2/15) dealing with the government crackdown on Buddhist monks, next to a number of articles on units (4/15), (5/15), and neighboring ones, covering the fighting and suffering during the Vietnam War.

A cluster of documents covering affairs in the Middle East is located in the lower right corner of the map around unit (10/15), next to a cluster on the Profumo–Keeler affair, a political scandal in Great Britain in the 1960s, on and around units (10/11) and (10/12). Above this area, on units (10/6) and neighboring ones we find articles on elections in Italy and possible coalitions, next to two units (10/3) and (10/4) covering elections in India. Similarly, all other units on the map can be identified to represent a topical cluster of news articles. For a more detailed discussion of the articles and topic clusters found on this map, we refer to [30].

While we find the SOM to provide a good topologically ordered representation of the various topics found in the article collection, no information about topical hierarchies can be identified from the resulting flat map. Apart from this, we find the size of the map to be quite large with respect to the number of topics identified. This is mainly because the size of the map has to be determined in advance, before any information about the number of topical clusters is available.

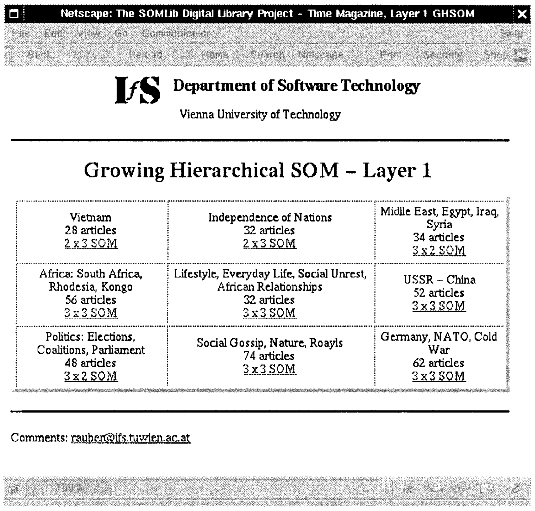
3) Hierarchical Archive of the TIME Magazine Collection: Based on the unit representing the mean of all data points at layer 0, the GHSOM training algorithm starts with a 2 × 2 SOM at the first layer. The training process for this map continues with additional units being added until the quantization error drops below a certain percentage of the overall quantization error of the unit at layer 0. The resulting first-layer map is depicted in Fig. 5. The map has grown twice, adding one row and one column, respectively, resulting in 3 × 3 units representing nine major topics in the document collection.

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

Based on this first separation of the most dominant topical clusters in the article collection, further maps were automatically trained to represent the various topics in more detail. This results in nine individual maps on the second layer, each representing the data of the respective higher layer unit in more detail. Some of the units on these second-layer maps were further expanded as distinct SOMs in the third layer.

The resulting second-layer maps are depicted in Fig. 6. Please note that the maps on the second layer have grown to different sizes according to the structure of the data. In particular, we find

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



Department of Software Technology

Vienna University of Technology

Growing Hierarchical SOM – Layer 1

Vietnam 28 articles 2 x 3 SOM	Independence of Nations 32 articles 2 x 3 SOM	Midlle East, Egypt, Iraq, Syria 34 articles 3 x 2 SOM
Africa: South Africa, Rhodesia, Kongo 56 articles 3 x 3 SOM	Lifestyle, Everyday Life, Social Unrest, African Relationships 32 articles 3 x 3 SOM	USSR – China 52 articles 3 x 3 SOM
Politics: Elections, Coalitions, Parliament 48 articles 3 x 2 SOM	Social Gossip, Nature, Roayls 74 articles 3 x 3 SOM	Germany, NATO, Cold War 62 articles 3 x 3 SOM

Comments: rauber@ifs.tuwien.ac.at

Fig. 5. *TIME* Magazine GHSOM Layer 1: 3 × 3 map of the *TIME* Magazine collection.

T148 T202 T519	*map:* T029 T051 T053 T228 T269 T313 T464 T470 T518 T545 T559
T322 T508	T434 T480 T498 T533
T363 T415	*map:* T320 T334 T390 T396 T414 T418

T066 T162 T282 T319	T094 T203 T204 T303 T335 T389 T443 T479
T296	T256 T270 T299
T111 T174 T285 T422 T536	*map:* T049 T098 T175 T188 T201 T225 T226 T227 T281 T284 T306

T026 T140 T253 T295	T060 T138 T333 T459	T068 T185 T446
map: T157 T163 T187 T215 T239 T252 T268 T283 T398 T509 T522	T059 T087 T104 T121 T294	*map:* T061 T067 T122 T267 T300 T361 T411

map: T021 T028 T048 T058 T065 T100 T358	T082 T238 T462 T477	T190 T255 T297 T408 T487 T524
T099 T109 T192 T224 T273	T272 T336 T407 T438	T057 T214 T237 T286 T289 T502 T549
T149 T151 T159 T302 T321 T403 T546 T547 T563	T110 T254 T391 T427 T442 T475 T525 T538	T123 T133 T160 T176 T221 T562

T034 T106 T345 T460 T490	T154 T444 T535	T478 T516
T301 T311 T399	T177 T276	T243 T280 T543
T050 T112 T262 T277 T382 T541 T556	T090 T279 T461 T537 T553	T081 T534 T544 T561

T032 T146 T153 T193 T275	T229 T244 T248 T347	T018 T042 T155 T156 T178 T194 T426 T539 T542
T101 T346 T351 T364 T365 T381 T394 T473	T062 T380 T392 T404	T135 T171 T245 T329 T495 T521
T025 T054 T353 T449 T472 T550 T557	T036 T257 T305 T491	T035 T134 T250 T425 T445

T040 T055 T063 T105 T196 T240 T205 T288 T341 T417 T497 T512 T526	T179 T476 T493 T523	*map:* T024 T070 T096 T130 T242 T304 T413 T424 T501 T560
T170 T315 T324 T337 T342 T354 T369 T529	T182 T198 T388 T400	T043 T107 T230 T247 T278 T331 T348 T485 T530

T030 T083 T119 T220 T222 T266 T326 T355	T120 T150 T180 T218 T219 T234 T309 T406 T431	T031 T235 T236 T274 T386 T330 T496 T503 T552
T033 T095 T191 T246 T551	T093 T152 T195 T249 T259 T528	T161 T189 T258 T298 T307 T350
T023 T069 T084 T097 T118 T136 T223 T370 T437 T514	T020 T091 T113 T184 T200 T310 T368 T412 T494	T064 T129 T147 T173 T217 T251 T312 T318 T356 T359 T430 T527

T102 T128 T143 T158 T172 T197 T261 T292 T357 T402	T019 T137 T233 T264 T383 T384 T507	T088 T108 T116 T145 T241 T263 T401 T513
T115 T131 T323 T463 T471 T504	T293 T436	T144 T186 T332 T367 T385 T405 T555
T045 T071 T072 T085 T086 T199 T231 T260	T017 T126 T183 T213 T287 T308 T317 T540	T047 T117 T181 T492 T511 T558

Fig. 6. *TIME* Magazine GHSOM Layer 2: 9 layer-2 maps of the *TIME* Magazine collection: 1 SOM per unit of the layer-1 map.

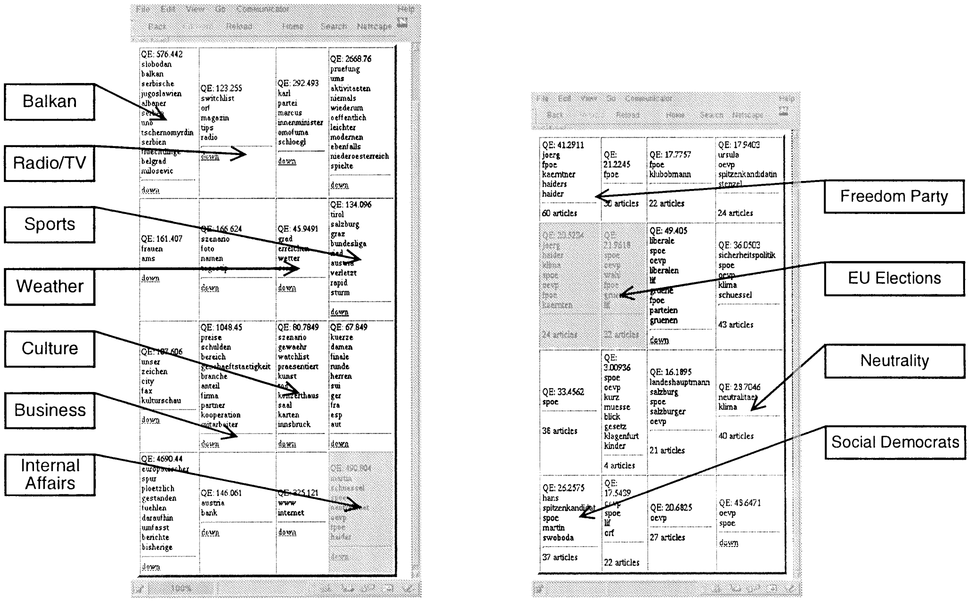
small 3 × 2 maps representing the articles of unit (3/1) and (1/3) of the first-layer map, and up to 3 × 3 maps for the units (1/2), (2/2), (3/2), (2/3), and (3/3). Taking a more detailed look at the first map of the second layer representing unit (1/1) on the first layer, we find it to give a clearer representation of articles covering the situation in Vietnam. Units (1/1) and (2/1) on this map represent articles on the fighting during the Vietnam War, whereas the remaining units represent articles on the internal conflict between the Catholic government and Buddhist monks. At this layer, the two units (2/1) and (2/3) have further been expanded to form separate maps with 3 × 3 units each at the third layer.

To give another example of the hierarchical structures identified during the GHSOM training process, we take a look at the 3 × 2 map representing the articles of unit (1/3) of the first-layer map. All of these articles were found to deal with political matters on the first layer. This common topic is now displayed in more detail at the resulting second-layer map. For example, we find unit (3/1) to represent articles on the elections in India. This unit is expanded to form a 3 × 4 map in the third layer. Next to these, on unit (3/2) we find articles covering the elections and discussions about political coalitions between Socialists and Christian Democrats in Italy. The remaining units on this map deal with different issues related to the Profumo–Keeler scandal in Great Britain, covering the political hearings in parliament, background information on this scandal and the persons involved, as well as related issues pertaining to the elections in Great Britain.

As a last example, consider the 3 × 3 map representing articles of unit (3/3) of the first layer. In the first layer, we find this unit to cover articles related to east–west relationships, mainly dealing with postwar Germany, the relationships between Germany and the Soviet Union and NATO. In the second-layer map, we find these topics to be separated clearer, with units (1/1) to (3/1) covering mainly Germany-related articles, whereas the other two topics are represented by the remaining four units. In this case, no further expansions onto third-layer maps were necessary, as all articles are already represented in sufficient detail on the units of the second-layer map.

4) Comparison of SOM and GHSOM Representation: When comparing the GHSOM with an SOM we can identify the locations of the articles on the nine second-layer maps on a corresponding 10 × 15 SOM. This allows us to view the hierarchical structure of the data on the flat map. We find that, for example, the cluster on Vietnam simply forms one larger coherent cluster on the flat map in the lower left corner of the map covering the rectangle spanned by the units (1/14) and (5/15). The same applies to the cluster of Middle-East affairs, which is represented by the map of unit (3/1) in the GHSOM. This cluster is mainly located in the lower right corner of the SOM. The cluster of political affairs, represented by unit (1/3) on the first layer of the GHSOM and explained in more detail on its subsequent layers, is spread across the right—hand side of the SOM, covering more or less all units on columns 9 and 10 and between rows 3 and 12. Note, that this common topic of political issues is not easily discernible from the overall map representation in the SOM, where exactly this hierarchical information is lost. The subdivision of this cluster on political matters becomes further evident when we consider the second-layer classification of this topic area, where the various subtopics are clearly separated, covering Indian elections, Italian coalitions, and the British Profumo–Keeler scandal.

Another interesting hierarchical structure not evident from the SOM is represented by the East–West relationships cluster



QE: 576.442 slobodan balkan serbische jugoslawien albaner serben uno tschernomyrdin serbien fluechtlinge belgrad milosevic down	QE: 123.255 switchlist orf magazin tips radio down	QE: 292.493 karl partei marcus innenminister omofuma schloegl down	QE: 2668.76 pruefung ums aktivitaeten niemals wiederum oeffentlich leichter modernen ebenfalls niederoesterreich spielte down
QE: 161.407 frauen ams down	QE: 166.624 szenario foto namen tagestip down	QE: 45.9491 grad erreichen wetter sonne down	QE: 134.096 tirol salzburg graz bundesliga ried austria verletzt rapid sturm down
QE: 107.606 unser zeichen city fax kulturschau down	QE: 1048.45 preise schulden bereich geschaeftstaetigkeit branche anteil firma partner kooperation mitarbeiter down	QE: 80.7849 szenario gewaehr watchlist praesentiert kunst tos konzerthaus saal karten innsbruck down	QE: 67.849 kuerze damen finale runde herren sui ger fra esp aut down
QE: 4690.44 europaeischer spur ploetzlich gestanden fuehlen daraufhin umfasst berichte bisherige down	QE: 146.061 austria bank down	QE: 325.121 www internet down	QE: 490.804 martin schuessel spoe neutralitaet oevp fpoe haider down

(a)

QE: 41.2911 joerg fpoe kaerntner haiders haider 60 articles	QE: 21.2245 fpoe 30 articles	QE: 17.7757 fpoe klubobmann 22 articles	QE: 17.9403 ursula oevp spitzenkandidatin stenzel 24 articles
QE: 23.5234 joerg haider klima spoe oevp fpoe kaernten 24 articles	QE: 21.9818 spoe oevp wahl fpoe gruene lif 22 articles	QE: 49.405 liberale spoe oevp liberalen lif gruene fpoe parteien gruenen down	QE: 36.0503 sicherheitspolitik spoe oevp klima schuessel 43 articles
QE: 33.4562 spoe 36 articles	QE: 3.00936 spoe oevp kurz muesse blick gesetz klagenfurt kinder 4 articles	QE: 16.1895 landeshauptmann salzburg spoe salzburger oevp 21 articles	QE: 28.7046 neutralitaet klima 40 articles
QE: 26.2575 hans spitzenkandidat spoe martin swoboda 37 articles	QE: 17.5439 oevp spoe lif orf 22 articles	QE: 20.6825 oevp 27 articles	QE: 48.6471 oevp spoe down

(b)

Fig. 7. (a) Top–layer map with 4×4 units showing general topics and (b) second-level maps with 4×4 units representing national politics.

on unit (3/3) of the GHSOM. When identifying the areas on the SOM that are represented by this branch in the GHSOM, we find that it covers two areas. This is, on the one hand, a group of seven units in the upper left corner of the SOM representing the Germany-related articles, whereas a second area in the upper right area of the SOM covers the NATO-related articles in this cluster. The relationship between these two subclusters is lost in the large SOM. This may be because of the size of the SOM, where the overall organization of the map needs to be determined during the very first training steps when the neighborhood range of the learning function still covers a large area of the SOM. A similar situation can be identified for several smaller clusters, which are scattered across different areas on the SOM, but nicely combined in the first layer of the GHSOM and further analyzed and separated as independent subclusters on subsequent layers.

Yet another interesting feature of the GHSOM we want to emphasize is the overall reduction in map size. During analysis, we found the second layer of the GHSOM to represent the data at about the same level of topical detail as the corresponding SOM. However, the number of units of all individual second-layer maps combined is only 69 as opposed to 150 units in the 10 × 15 SOM. With the GHSOM model, this number of units is determined automatically, and only the necessary number of units is created for each level of detail representation required by the respective layer. Furthermore, not all branches are grown to the same depth of the hierarchy. As can be seen from Fig. 6, only some of the units are further expanded in a third-layer map. With the resulting maps at all layers of the hierarchy being rather small, activation calculation and winner evaluation of the GHSOM is by orders of magnitude faster than in the SOM model. Apart from the speed-up gained by the reduced network size, orientation for the user is highly improved as compared to the rather huge maps which cannot be easily comprehended as a whole.

B. Experiment 2: Two Hierarchies of Newspaper Articles From "Der Standard"

1) Data Representation: In the second experiment, we will take a closer look at the influence of parameter τ_1, providing a tradeoff between shallow and deep hierarchies, as well as the topology-preserving orientation of the various maps in different branches of the hierarchy. For these, we use a larger collection of 11 627 articles from the Austrian newspaper *Der Standard* covering the second quarter of 1999. To be used for map training, a vector-space representation of the single documents is created by full-text indexing. Instead of defining language or content specific stop word lists, we discard terms that appear in more than 813 (7%) or in less than 65 articles (0.56%). We end up with a vector dimensionality of 3799 unique terms. Thus, the 11 627 articles are represented by automatically extracted 3799-dimensional feature vectors of word histograms, weighted by a $tf \times idf$ weighting scheme and normalized to unit length. These feature vectors are used to train two GHSOMs. The mqe_0 of the layer-0 map evaluates to 12 180.3, serving as the basis for the global stopping criterion.

2) Deep Hierarchy: Training the GHSOM with parameters $\tau_1 = 0.07$ and $\tau_2 = 0.0035$ results in a rather deep hierarchical structure of up to 13 layers. Since it is impossible to present the complete topic hierarchy of three months of news articles, we will concentrate on some sample topical sections. The first-layer map depicted in Fig. 7(a) has grown to a size of 4 × 4 units, all of which are expanded at subsequent layers. Among the well separated main topical branches we find sports, culture, radio- and TV programs, the political situation in the Balkans, national politics, business, or weather reports, to name a few. These topics are clearly identifiable by the automatically extracted keywords using the LabelSOM technique [17], [18], such as *weather*, *sun*, *reach*, *degrees* for the section on weather reports.[3] The branch of articles covering the political

[3] We provide English translations for the original German labels.

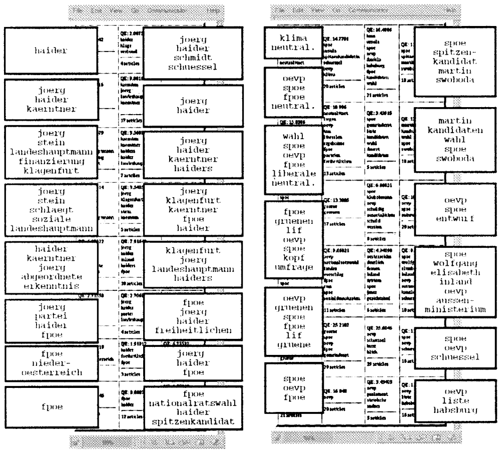

haider
joerg haider schmidt schuessel
joerg haider kaerntner
joerg haider
joerg stein landeshauptmann finanzierung klagenfurt
joerg haider kaerntner haiders
joerg stein schlaegt soziale landeshauptmann
joerg klagenfurt kaerntner fpoe haider
haider kaerntner joerg abgeordnete erkenntnis
klagenfurt joerg landeshauptmann haiders
joerg partei haider fpoe
fpoe joerg haider freiheitlichen
fpoe nieder-oesterreich
joerg haider fpoe
fpoe
fpoe nationalratswahl haider spitzenkandidat
klima neutral.
spoe spitzen-kandidat martin swoboda
oevp spoe fpoe neutral.
wahl spoe oevp fpoe liberale neutral.
martin kandidaten wahl spoe swoboda
fpoe gruenen lif oevp spoe kopf umfrage
oevp spoe entwurf
spoe wolfgang elisabeth inland oevp aussen-ministerium
oevp gruenen spoe fpoe lif gruene
spoe oevp schuessel
spoe oevp fpoe
oevp liste habsburg

Fig. 8. Two neighboring second-layer maps on national politics.

situation on the Balkan is located in the upper left corner of the top-layer map labeled with *Balkan*, *Slobodan Milosevic*, *Serbs*, *Albanians*, *UNO*, *Refugees*, and others.

We find the branch on national politics in the lower right corner of this map listing the three largest political parties of Austria as well as two key politicians as labels. This unit has been expanded to form a 4 × 4 map in the second layer as shown in Fig. 7(b). The upper left area of this map is dominated by articles related to the Freedom Party, whereas, for example, articles focusing on the Social Democrats are located in the lower left corner. Other dominant clusters on this map are neutrality, or the elections to the European parliament, with one unit carrying specifically the five political parties as well as the term *election* as labels. Two units of this second-layer map are further expanded in a third layer, such as, for example, the unit in the lower right corner representing articles related to the coalition of the People's Party and the Social Democrats. These articles are represented in more detail by a 3 × 4 map in the third layer.

3) Shallow Hierarchy: To show the effects of different parameter settings we trained a second GHSOM with τ_1 set to half of the previous value ($\tau_1 = 0.035$), while τ_2, i.e., the absolute granularity of data representation, remained unchanged. This leads to a more shallow hierarchical structure of only up to seven layers, with the first-layer map evolving to a size of 7 × 4. Again, we find the most dominant branches to be, for example, sports, located in the upper right corner of the map, national politics in the lower right corner, Internet-related articles on the left-hand side of the map, to name a few. However, because of the large size of the resulting first-layer map, a fine-grained representation of the data is already provided at this layer. This results in some larger clusters being represented by two neighboring units already at the first layer, rather than being split up at a lower layer of the hierarchy. For example, we find the cluster on national politics to be represented by two neighboring units. One of these, on position (6/4), covers solely articles related to the Freedom Party and its political leader Jörg Haider, representing one of the most dominant political topics in Austria for some time now, resulting in an accordingly large number of news articles covering this topic. The neighboring unit to the right, i.e., located in the lower right corner on position (7/4), covers other aspects of national politics, with one of the main topics being the elections to the European parliament. Fig. 8 shows these two second-layer maps.

However, we also find articles related to the Freedom Party on this second branch, covering the more general national politics, reporting on their role and campaigns for the elections to the European parliament. As might be expected these are closely related to the other articles on the Freedom Party, which are located in the neighboring branch to the left. Obviously, we would like them to be presented on the left-hand side of this map, so as to allow the transition from one map to the next, with a continuous orientation of topics. Because of the initialization of the added maps during the training process, this continuous orientation is preserved, as can easily be seen from the automatically extracted labels provided in Fig. 8. Continuing from the second-layer map of unit (6/4) to the right, we reach the according second-layer map of unit (7/4), where we first find articles focusing on the Freedom Party, before moving on to the Social Democrats, the People's Party, the Green Party, and the Liberal Party. As all units at layer two in these branches have a qe below 42.63, no unit is further expanded at a third layer.

We thus find the global orientation to be well preserved in this map. Even though the cluster of national politics is split into two dominant subclusters in the more shallow hierarchy, the articles are organized correctly on the two separate maps in the second layer of the hierarchy. This allows the user to continue his or her exploration across map boundaries. For this purpose, the labels of the upper layer's neighboring unit serves a general guideline as to which topic is covered by the neighboring map. In the deeper hierarchy, these two subclusters are represented within one single branch in the second layer, covering the upper and the lower area of the map, respectively.

VI. Conclusion

In this paper, we have described the GHSOM. The characteristic feature of this novel neural-network model is its adaptive architecture which grows during its unsupervised training process to uncover the hierarchical structure of the analyzed data collection. In a nutshell, the GHSOM has a layered architecture composed of independent SOM-like neural networks. These networks determine their structure during the training process as well.

The major benefits of our GHSOM model compared with the standard SOM are the following. First, the overall training time is largely reduced since only the necessary number of units are developed to organize the data collection at a certain degree of detail. Second, the GHSOM uncovers the hierarchical structure of the data by its very architecture, thus allowing the user to understand and analyze large amounts of data in an explorative way. Third, with the various emerging maps at each layer of the hierarchy being rather small in size, it is much easier for the user to keep an overview of the various clusters. Last, but not least, by ensuring a consistent global orientation of the individual maps in the respective layers, the topological similarities of neighboring maps are preserved. Thus, navigation across map boundaries is facilitated, allowing the exploration of similar clusters that are represented by neighboring branches in the GHSOM structure.

We have shown the potential of the GHSOM with an information retrieval application, namely the topical clustering of free-form documents. This, by its very nature, is a challenging application for clustering algorithms because of the high-dimensional and noisy feature spaces. The results of the experiments clearly indicated that the GHSOM successfully identified the topical clusters of the document collections.

References

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, Sept. 1999.

[2] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, pp. 59–69, 1982.

[3] S. Kaski, J. Kangas, and T. Kohonen, "Bibliography of self-organizing map (SOM) papers 1981–1997," *Neural Comput. Surveys*, vol. 1, no. 3 & 4, pp. 1–176, 1998.

[4] J. Blackmore and R. Miikkulainen, "Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map," in *Proc. IEEE Int. Conf. Neural Networks (ICNN'93)*, vol. 1, San Francisco, CA, 1993, pp. 450–455.

[5] B. Fritzke, "Growing grid—A self-organizing network with constant neighborhood range and adaption strength," *Neural Processing Lett.*, vol. 2, no. 5, pp. 1–5, 1995.

[6] D. Alahakoon, S. K. Halgamuge, and B. Srinivasan, "Dynamic self-organizing maps with controlled growth for knowledge discovery," *IEEE Trans. Neural Networks*, vol. 11, pp. 601–614, May 2000.

[7] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1989.

[8] J. A. Kangas, T. Kohonen, and J. T. Laaksonen, "Variants of self-organizing maps," *IEEE Trans. Neural Networks*, vol. 1, pp. 93–99, Mar. 1990.

[9] T. Kohonen, *Self-Organizing Maps*. Berlin, Germany: Springer-Verlag, 1995.

[10] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela, "Self-organization of a massive document collection," *IEEE Trans. Neural Networks*, vol. 11, pp. 574–585, May 2000.

[11] X. Lin, "A self-organizing semantic map for information retrieval," in *Proc. 14th Annu. Int. ACM SIGIR Conf. Res. Development Information Retrieval (SIGIR91)*, Chicago, IL, Oct. 13–16, 1991, ACM, pp. 262–269.

[12] D. Merkl, "Text classification with self-organizing maps: Some lessons learned," *Neurocomput.*, vol. 21, no. 1–3, pp. 61–77, 1998.

[13] D. Merkl and A. Rauber, "Document classification with unsupervised neural networks," in *Soft Computing in Information Retrieval*, F. Crestani and G. Pasi, Eds: Physica-Verlag, 2000, pp. 102–121.

[14] D. G. Roussinov and H. Chen, "Information navigation on the web by clustering and summarizing query results," *Inform. Processing Manage.*, vol. 37, pp. 789–816, 2001.

[15] A. Ultsch, "Self-organizing neural networks for visualization and classification," in *Information and Classification: Concepts, Methods, and Applications*, O. Opitz, B. Lausen, and R. Klar, Eds. Dortmund, Germany: Springer-Verlag, 1992, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 307–313.

[16] D. Merkl and A. Rauber, "Alternative ways for cluster visualization in self-organizing maps," in *Proc. Workshop Self-Organizing Maps (WSOM97)*, T. Kohonen, Ed. Espoo, Finland: Helsinki Univ. Technol., HUT, June 4–6, 1997, pp. 106–111.

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

[18] D. Merkl and A. Rauber, "Automatic labeling of self-organizing maps for information retrieval," in *Proc. 6th Int. Conf. Neural Inform. Processing (ICONIP99)*, Perth, Australia, Nov. 16–20, 1999.

[19] S. Santini, "The self-organizing field," *IEEE Trans. Neural Networks*, vol. 7, pp. 1415–1423, Nov. 1996.

[20] C. M. Bishop, M. Svensen, and C. K. I. Williams, "GTM: The generative topographic mapping," *Neural Comput.*, vol. 10, no. 1, pp. 215–235, 1998.

[21] R. Miikkulainen, "Script recognition with hierarchical feature maps," *Connection Sci.*, vol. 2, pp. 83–101, 1990.

[22] P. Koikkalainen and E. Oja, "Self-organizing hierarchical feature maps," in *Proc. Int. Joint Conf. Neural Networks*, vol. 2, San Diego, CA, 1990, pp. 279–284.

[23] P. Koikkalainen, "Fast deterministic self-organizing maps," in *Proc. Int. Conf. Artificial Neural Networks*, vol. 2, Paris, France, 1995, pp. 63–68.

[24] H.-U. Bauer and T. Villmann, "Growing a hypercubical output space in a self-organizing feature map," *IEEE Trans. Neural Networks*, vol. 8, pp. 226–233, Mar. 1997.

[25] M. Dittenbach, D. Merkl, and A. Rauber, "The growing hierarchical self-organizing map," in *Proc. Int. Joint Conf. Neural Networks (IJCNN 2000)*, S. Amari, C. L. Giles, M. Gori, and V. Puri, Eds. Como, Italy: IEEE Comput. Soc., July 24–27, 2000, vol. VI, pp. 15–19.

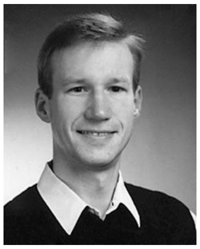
[26] M. Dittenbach, A. Rauber, and D. Merkl, "Recent advances with the growing hierarchical self-organizing map," in *Proc. 3rd Workshop Self-Organizing Maps*, Advances in Self-Organizing Maps, N. Allinson, H. Yin, L. Allinson, and J. Slack, Eds., Lincoln, U.K., June 13–15, 2001, pp. 140–145.

[27] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Inform. Processing Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[28] H. R. Turtle and W. B. Croft, "A comparison of text retrieval models," *Comput. J.*, vol. 35, no. 3, pp. 279–290, 1992.

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

[30] A. Rauber and D. Merkl, "Using self-organizing maps to organize document collections and to characterize subject matters: How to make a map tell the news of the world," in *Proc. 10th Int. Conf. Database Expert Syst. Applicat. (DEXA99)*, T. Bench-Capon, G. Soda, and A. M. Tjoa, Eds. Florence, Italy, Sept. 1–3, 1999, Lecture Notes in Computer Science 1677, pp. 302–311.



Andreas Rauber (S'93–M'01) received the M.Sc. and Ph.D. degrees in computer science from the Vienna University of Technology, Vienna, Austria, in 1997 and 2000, respectively.

Since 1997, he has been a Member of the Academic Faculty at the Department of Software Technology at the Vienna University of Technology. In 2002, he held an ERCIM Research Fellowship at the Italian National Research Council (CNR), Pisa, Italy. He is currently an ERCIM Research Fellow at INRIA, Paris, France. He has published more than 30 papers in refereed journals and international conferences. His current research interests include neural computation, digital libraries, and information visualization.

Dr. Rauber received the OeGAI Award of the Austrian Society for Artificial Intelligence in 1998.

Dieter Merkl received the Diploma and Doctoral degrees in social and economic sciences from the University of Vienna, Vienna, Austria, in 1989 and 1995, respectively.

He is Associate Professor with the Department of Software Technology, Vienna University of Technology, Vienna, Austria. From 1990 to 1994, he held a research position at the University of Vienna. Since 1995, he has been a Member of the Academic Faculty at Vienna University of Technology. During 1997, he was Visiting Research Fellow with the Department of Computer Science, Royal Melbourne Institute of Technology, Melbourne, Australia. He has published more than 80 articles in refereed journals and international conferences. His current research interests include neural computation, information retrieval, and software engineering.

Dr. Merkl is an affiliate member of the IEEE Computer Society.

Michael Dittenbach received the Diploma degree in computer science from the Vienna University of Technology, Vienna, Austria.

He is currently a Research Assistant with the E-Commerce Competence Center EC3, Adaptive Multilingual Interfaces Group, and a Junior Researcher with the Department of Software Technology, Vienna University of Technology. He has published several papers at international conferences. Currently, his main research interests include natural language processing, cross-language information retrieval, text mining, digital libraries, and neurocomputing.