

# Business, Culture, Politics, and Sports – How to Find Your Way Through a Bulk of News? On Content-Based Hierarchical Structuring and Organization of Large Document Archives

Michael Dittenbach<sup>1</sup>, Andreas Rauber<sup>2</sup>, Dieter Merkl<sup>2</sup>

<sup>1</sup> E-Commerce Competence Center – EC3,  
Siebensterngasse 21/3, A-1070 Wien, Austria

<sup>2</sup> Institut für Softwaretechnik, Technische Universität Wien,  
Favoritenstraße 9–11/188, A-1040 Wien, Austria  
[www.ifs.tuwien.ac.at/~mbach, ~andi, ~dieter](http://www.ifs.tuwien.ac.at/~mbach, ~andi, ~dieter)

**Abstract.** With the increasing amount of information available in electronic document collections, methods for organizing these collections to allow topic-oriented browsing and orientation gain increasing importance. The *SOMLib* digital library system provides such an organization based on the *Self-Organizing Map*, a popular neural network model by producing a map of the document space. However, hierarchical relations between documents are hidden in the display. Moreover, with increasing size of document archives the required maps grow larger, thus leading to problems for the user in finding proper orientation within the map. In this case, a hierarchically structured representation of the document space would be highly preferable. In this paper, we present the *Growing Hierarchical Self-Organizing Map*, a dynamically growing neural network model, providing a content-based hierarchical decomposition and organization of document spaces. This architecture evolves into a hierarchical structure according to the requisites of the input data during an unsupervised training process. A recent enhancement of the training process further ensures proper orientation of the various topical partitions. This facilitates intuitive navigation between neighboring topical branches. The benefits of this approach are shown by organizing a real-world document collection according to semantic similarities.

## 1 Introduction

With the increasing amount of textual information stored in digital libraries, means to organize and structure this information have gained importance. Specifically an organization by content, allowing topic-oriented browsing of text collections, provides a highly intuitive approach to exploring document collections. As one of the most successful methods applied in this field we find the *Self-Organizing Map (SOM)* [5], a popular unsupervised neural network model, which is frequently being used to provide a map-based representation of document archives [7, 2, 10, 6]. In such a representation, documents on similar topics are located next to each other. The

obvious benefit for the user is that navigation in the document archive is similar to the well-known task of navigating in a geographical map. With the *SOMLib* digital library [9] we developed a system using the *SOM* as its core module to provide content-based access to document archives. This allows the user to obtain an overview of the topics covered in a collection, and their importance with respect to the amount of information present in each topical section.

While these characteristics made the *SOM* a prominent tool for organizing document collections, most of the research work aims at providing one single map representation for the complete document archive. As a consequence, hierarchical relations between documents are lost in the display. Moreover, it is only natural that with increasing size of the document archive the maps for representing the archive grow larger, thus leading to problems for the user in finding proper orientation within the map. We believe that the representation of hierarchical document relations is vital for the usefulness of map-based document archive visualization approaches.

In this paper we argue in favor of establishing such a hierarchical organization of the document space based on a novel neural network architecture, the *Growing Hierarchical Self-Organizing Map (GHSOM)* [3]. The distinctive feature of this model is its problem dependent architecture which develops during the unsupervised training process. Starting from a rather small high-level *SOM*, which provides a coarse overview of the various topics present in a document collection, subsequent layers are added where necessary to display a finer subdivision of topics. Each map in turn grows in size until it represents its topic to a sufficient degree of granularity. Since usually not all topics are present equally strong in a collection, this leads to an unbalanced hierarchy, assigning more “map-space” to topics that are more prominent in a given collection. This allows the user to approach and intuitively browse a document collection in a way similar to conventional libraries.

The hierarchical structuring imposed on the data represents a rather strong separation of clusters mapped onto different branches. While this is a highly desirable characteristic helping in understanding the topical cluster structure in large data sets, it may lead to misinterpretations when long-streched clusters are mapped and expanded on two neighboring, yet different units of the *SOM*. This can be alleviated by ensuring proper orientation of the maps in the various branches of the hierarchy, allowing navigation between branches. We present the the benefits of such a hierarchical organization of digital libraries, as well as the stability of the process using a set of experiments based on a collection of newspaper articles from the daily Austrian newspaper *Der Standard*. Specifically, we compare two different representations of the topical hierarchy of this archive resulting from different parameter settings.

The remainder of this paper is organized as follows. In Section 2 we provide a brief review of related architectures followed by a description of the principles of the *SOM* and *GHSOM* training in Section 3. Subsequently, we provide a detailed discussion of our experimental results in Section 4 as well as some conclusions in Section 5.

## 2 Related Work

A number of extensions and modifications have been proposed over the years in order to enhance the applicability of *SOMs* to data mining, specifically inter- and intra-cluster similarity identification. The *Hierarchical Feature Map (HFM)* [8] addresses the problem of hierarchical data representation 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 due to the fact that the architecture of the network has to be defined in advance, i.e. the number of layers and the size of the maps at each layer is fixed prior to network training. This leads to the definition of a balanced tree which is used to represent the data. What we want, however, is a network architecture definition based on the actual data presented to the network.

The shortcoming of having to define the size of the *SOM* in advance has been addressed in several models, such as the *Incremental Grid Growing (IGG)* [1] or *Growing Grid (GG)* [4] models. The former allows the adding of new units at the boundary of the map, while connections within the map may be removed according to some threshold settings, possibly resulting in several separated, irregular map structures. The latter model, on the other hand, adds rows and columns of units during the training process, starting with an initial  $2 \times 2$  *SOM*. This way the rectangular layout of the *SOM* grid is preserved.

## 3 Content-Based Organization of Text Archives

### 3.1 Feature Extraction

In order to allow content-based classification of documents we need to obtain a representation of their content. One of the most common representations uses word frequency counts based on full text indexing. A list of all words present in a document collection is created to span the feature space within which the documents are represented. While hand-crafted stop word lists allow for specific exclusion of frequently used words, statistical measures may be used to serve the same purpose in a more automatic way. For our experiments we thus remove all words that appear either in too many documents within a collection (e.g. say in more than 50% of all documents) or in too few (say, less than 5 documents) as these words do not contribute to content representation. The words are further weighted according to the standard  $tf \times idf$ , i.e. term frequency times inverse document frequency, weighting scheme [11]. This weighting scheme assigns high values to words that are considered important for content representation. The resulting feature vectors may further be used for *SOM* training.

### 3.2 Self-Organizing Map

The *Self-Organizing Map* is an unsupervised neural network providing a mapping from a high-dimensional input space to a usually two-dimensional output space

while preserving topological relations as faithfully as possible. The *SOM* consists of a set of  $i$  units arranged in a two-dimensional grid, with a weight vector  $m_i \in \mathbb{R}^n$  attached to each unit. Elements from the high dimensional input space, referred to as input vectors  $x \in \mathbb{R}^n$ , are presented to the *SOM* and the activation of each unit for the presented input vector is calculated using an activation function. Commonly, the Euclidean distance between the weight vector of the unit and the input vector serves as the activation function. In the next step the weight vector of the unit showing the highest activation (i.e. the smallest Euclidean distance) is selected as the ‘winner’ and is modified as to more closely resemble the presented input vector. Pragmatically speaking, the weight vector of the winner is moved towards the presented input signal by a certain fraction of the Euclidean distance as indicated by a time-decreasing learning rate  $\alpha$ . Thus, this unit’s activation will be even higher the next time the same input signal is presented. Furthermore, the weight vectors of units in the neighborhood of the winner as described by a time-decreasing neighborhood function  $\epsilon$  are modified accordingly, yet to a less strong amount as compared to the winner. This learning procedure finally leads to a topologically ordered mapping of the presented input signals. Similar input data is mapped onto neighboring regions on the map.

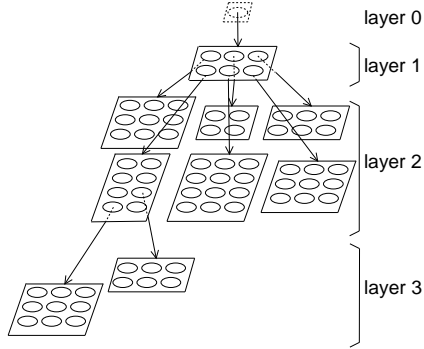
### 3.3 Growing Hierarchical Self-Organizing Map

The key idea of the *GHSOM* is to use a hierarchical structure of multiple layers where each layer consists of a number of independent *SOMs*. One *SOM* is used at the first layer of the hierarchy. For every unit in this map a *SOM* might be added to the next layer of the hierarchy. This principle is repeated with the third and any further layers of the *GHSOM*.

Since one of the shortcomings of *SOM* usage is its fixed network architecture we rather use an incrementally growing version of the *SOM*. This relieves us from the burden of predefining the network’s size, which is rather determined during the unsupervised training process. We start with a layer 0, which consists of only one single unit. The weight vector of this unit is initialized as the average of all input data. The training process then basically starts with a small map of  $2 \times 2$  units in layer 1, which is self-organized according to the standard *SOM* training algorithm.

This training process is repeated for a fixed number  $\lambda$  of training iterations. Ever after  $\lambda$  training iterations the unit with the largest deviation between its weight vector and the input vectors represented by this very unit is selected as the error unit. In between the error unit and its most dissimilar neighbor in terms of the input space either a new row or a new column of units is inserted. The weight vectors of these new units are initialized as the average of their neighbors.

An obvious criterion to guide the training process is the quantization error  $q_i$ , calculated as the sum of the distances between the weight vector of a unit  $i$  and the input vectors mapped onto this unit. It is used to evaluate the mapping quality of a *SOM* based on the mean quantization error (*MQE*) of all units in the map. A map grows until its *MQE* is reduced to a certain fraction  $\tau_1$  of the  $q_i$  of the unit  $i$  in the preceding layer of the hierarchy. Thus, the map now represents the data mapped onto the higher layer unit  $i$  in more detail.



**Fig. 1.** *GHSOM* reflecting the hierarchical structure of the input data.

As outlined above the initial architecture of the *GHSOM* consists of one *SOM*. This architecture is expanded by another layer in case of dissimilar input data being mapped on a particular unit. These units are identified by a rather high quantization error  $q_i$  which is above a threshold  $\tau_2$ . This threshold basically indicates the desired granularity level of data representation as a fraction of the initial quantization error at layer 0. In such a case, a new map will be added to the hierarchy and the input data mapped on the respective higher layer unit are self-organized in this new map, which again grows until its *MQE* is reduced to a fraction  $\tau_1$  of the respective higher layer unit's quantization error  $q_i$ .

A graphical representation of a *GHSOM* is given in Figure 1. The map in layer 1 consists of  $3 \times 2$  units and provides a rough organization of the main clusters in the input data. The six independent maps in the second layer offer a more detailed view on the data. Two units from one of the second layer maps have further been expanded into third-layer maps to provide sufficiently granular data representation.

Depending on the desired fraction  $\tau_1$  of *MQE* reduction we may end up with either a very deep hierarchy with small maps, a flat structure with large maps, or – in the most extreme case – only one large map, which is similar to the *Growing Grid*. The growth of the hierarchy is terminated when no further units are available for expansion. It should be noted that the training process does not necessarily lead to a balanced hierarchy in terms of all branches having the same depth. 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 create deeper branches than others.

The growth process of the *GHSOM* is mainly guided by the two parameters  $\tau_1$  and  $\tau_2$ , which merit further consideration.

- $\tau_2$ : Parameter  $\tau_2$  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. 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*.

- $\tau_1$ : This parameter controls the actual growth process of the *GHSOM*. Basically, hierarchical data can be represented in different ways, favoring either (a) lower hierarchies with rather detailed refinements presented at each subsequent layer, or (b) deeper 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. 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.

Thus, the smaller the parameter  $\tau_1$ , 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  $\tau_1$  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  $\tau_1$ , the more shallow the hierarchy, and that, the lower the setting of parameter  $\tau_2$ , the larger the number of units in the resulting *GHSOM* network will be.

In order 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 parents' maps. This can be achieved by creating a coherent initialization of the units of a newly created map, i.e. by adding a fraction of the weight vectors in the neighborhood of the parent unit. This initial orientation of the map is preserved during the training process. By providing a global orientation of all maps in the hierarchy, potentially negative effects of splitting a large cluster into two neighboring branches can be alleviated, as it is possible to navigate across map boundaries to neighboring maps.

## 4 Two Hierarchies of Newspaper Articles

For the experiments presented hereafter we use a collection of 11,627 articles from the Austrian daily 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 rather 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 3,799 unique terms. The 11,627 articles thus are represented by automatically extracted 3,799-dimensional feature vectors of word histograms weighted by a  $tf \times idf$  weighting scheme and normalized to unit length.

#### 4.1 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.<sup>1</sup> The layer 1 map depicted in Figure 2(a) grows to a size of  $4 \times 4$  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 on the Balkan, *Internal Affairs*, *Business*, or *Weather Reports*, to name but a few. These topics are clearly identifiable by the automatically extracted keywords using the *LabelSOM* technique [10], such as *weather*, *sun*, *reach*, *degrees* for the section on *Weather Reports*<sup>2</sup>. The branch of articles covering the political 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.

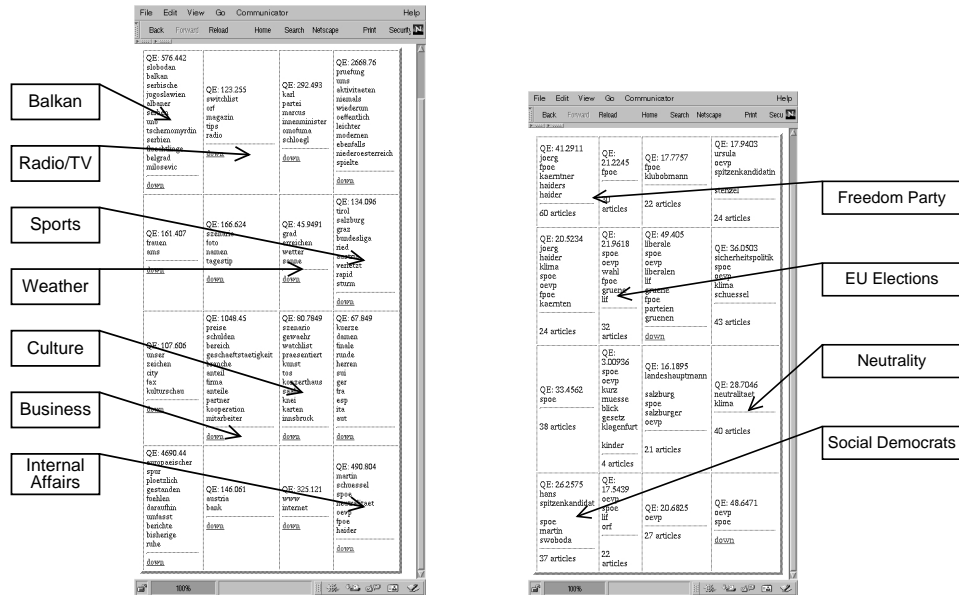
We find the branch on *Internal Affairs* in the lower right corner of this map listing the three largest political parties of Austria as well as two key politicians as labels. This unit has been expanded to form a  $4 \times 4$  map in the second layer as shown in Figure 2(b). The upper left area of 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  $3 \times 4$  map in the third layer.

#### 4.2 Shallow Hierarchy

To show the effects of different parameter settings we trained a second *GHSOM* with  $\tau_1$  set to half of the previous value ( $\tau_1 = 0.035$ ), while  $\tau_2$ , i.e. the absolute granularity of data representation, remained unchanged. This leads to a more shallow hierarchical structure of only up to 7 layers, with the layer 1 map growing to a size of  $7 \times 4$  units. Again, we find the most dominant branches to be, for example, *Sports*, located in the upper right corner of the map, *Internal Affairs* in the lower right corner, *Internet*-related articles on the left hand side of the map, to name but a few. However, due to the large size of the resulting first layer map, a fine-grained

<sup>1</sup> The maps are available for interactive exploration at [http://www.ifs.tuwien.ac.at/~andi/somlib/experiments\\_standard](http://www.ifs.tuwien.ac.at/~andi/somlib/experiments_standard)

<sup>2</sup> We provide English translations for the original German labels.



(a) Top layer map:  $4 \times 4$  units; Main topics (b) Second layer map:  $4 \times 4$  units; Internal Affairs

Fig. 2. Top and second level map.

representation of the data is already provided at this layer. This results in some larger clusters to be represented by two neighboring units already at the first layer, rather than being split up in a lower layer of the hierarchy. For example, we find the cluster on *Internal Affairs* 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 *Internal Affairs*, with one of the main topics being the elections to the *European Parliament*. Figure 3 shows these two second-layer maps.

However, we also find, articles related to the *Freedom Party* on this second branch covering the more general *Internal Affairs*, 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. Due to 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 Figure 3. Continuing

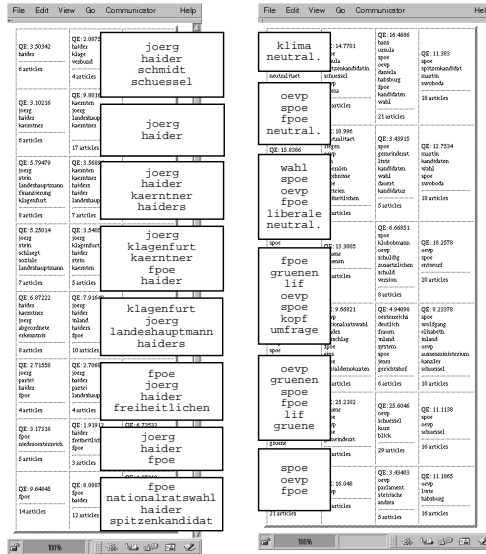


Fig. 3. Two neighboring second-layer maps on Internal Affairs

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*.

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

## 5 Conclusions

Automatic topical organization is crucial for providing intuitive means of exploring unknown document collections. While the *SOM* has proven capable of handling the complexities of content-based document organization, its applicability is limited, firstly, by the size of the resulting map, as well as secondly, by the fact that hierarchical relations between documents are lost within the map display.

In this paper we have argued in favour of a hierarchical representation of document archives. Such an organization provides a more intuitive means for exploring and understanding large information spaces. The *Growing Hierarchical Self-Organizing Map (GHSOM)* has shown to provide this kind of representation by adapting both its hierarchical structure as well as the sizes of each individual map

to represent data at desired levels of granularity. It fits its architecture according to the requirements of the input space, relieving the user from having to define a static organization prior to the training process.

Multiple experiments have shown both its capabilities of hierarchically organizing document collection according to their topics, as well as the benefits of providing a better overview of, especially, larger collections, where single map-based representations tend to become unacceptably large. Furthermore, by preserving a global orientation of the individual maps, navigation between neighboring maps is facilitated. The presented model thus allows the user to intuitively explore an unknown document collection by browsing through the topical sections.

## References

1. J. Blackmore and R. Miikkulainen. Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map. In *Proceedings of the IEEE International Conference on Neural Networks (ICNN'93)*, volume 1, pages 450–455, San Francisco, CA, USA, 1993. <http://ieeexplore.ieee.org/>.
2. H. Chen, C. Schuffels, and R. Orwig. Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation*, 7(1):88–102, 1996. <http://ai.BPA.arizona.edu/papers/>.
3. M. Dittenbach, D. Merkl, and A. Rauber. The growing hierarchical self-organizing map. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2000)*, volume VI, pages 15 – 19, Como, Italy, 2000. IEEE Computer Society. <http://www.ifs.tuwien.ac.at/ifs/research/publications.html>.
4. B. Fritzke. Growing Grid – A self-organizing network with constant neighborhood range and adaption strength. *Neural Processing Letters*, 2(5):1 – 5, 1995. <http://pikas.inf.tu-dresden.de/~fritzke>.
5. T. Kohonen. *Self-organizing maps*. Springer-Verlag, Berlin, 1995.
6. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self-organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000. <http://ieeexplore.ieee.org/>.
7. X. Lin. A self-organizing semantic map for information retrieval. In *Proceedings of the 14. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR91)*, pages 262–269, Chicago, IL, October 13 - 16 1991. ACM. <http://www.acm.org/dl>.
8. R. Miikkulainen. Script recognition with hierarchical feature maps. *Connection Science*, 2:83 – 101, 1990.
9. A. Rauber and D. Merkl. The SOMLib Digital Library System. In *Proceedings of the 3. European Conference on Research and Advanced Technology for Digital Libraries (ECDL99)*, LNCS 1696, pages 323–342, Paris, France, 1999. Springer. <http://www.ifs.tuwien.ac.at/ifs/research/publications.html>.
10. 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 *Proceedings of the 10. International Conference on Database and Expert Systems Applications (DEXA99)*, LNCS 1677, pages 302–311, Florence, Italy, 1999. Springer. <http://www.ifs.tuwien.ac.at/ifs/research/publications.html>.
11. G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

Michael Dittenbach, Andreas Rauber, and Dieter Merkl: Business, Culture, Politics, and Sports – How to Find Your Way Through a Bulk of News? On Content-Based Hierarchical Structuring and Organization of Large Document Archives  
In: Proceedings of the 12th International Conference on Database and Expert Systems Applications, Springer Lecture Notes in Computer Science, Sept. 3-7 2001, Munich, Germany, Springer, 2001.