Semantic Navigation Maps for Information Agents in Environment Information Systems

Wolfgang Benn[1], Günther Beyrich[1] und Otmar Görlitz[1]

Abstract

The automated retrieval of information in Environment Information Systems by information agents is severely hindered by the heterogeneity of these systems. For the decision if an information is a relevant answer to a query, information agents need to understand semantics and context of the query. In this paper we propose our concept of describing and exploiting contextual relations in the domain of environment information on the base of Kohonen's self-organizing feature maps. We apply this concept for the automated semantic classification of query results into an existing knowledge space. Using this technique, information agents can be endowed with domain knowledge. Thus they are enabled to retrieve information in heterogeneous Environment Information Systems matching the context of the query.

1 Introduction

1.1 Heterogeneous Environment Information Systems

The growing public interest in environment information and the availability of the internet as broad and easy to use information source increased the request to make environment data available via this medium to the public. Moreover, the European Council's guideline on public availability of environment data secures the distinct right of information for citizens. The federal states of Germany are establishing so called Environment Information Systems to meet this guideline.

However, these Environment Information Systems are often created parallel and independently. This inevitably leads to very heterogeneous structures and content of these systems. Additionally, many ecological research projects also make their results available in the World-Wide-Web. They too use their own schema for structuring the information which increases the general incompatibility of the information sources.

[1] Chemnitz University of Technology, Department of Computer Science, D-09107 Chemnitz, Germany

The general heterogeneity causes a new problem for the user. Although much information is available, for the query on a certain part of environment data, he needs to know first, which information system stores relevant data. To support the search for appropriate information sources, the federal environment office recommends the Environment Data Catalogue (in German: UDK). The UDK is a meta-information system which stores information on available environment data and the Environment Information Systems where these data can be found. Furthermore, the UDK contains an extensive thesaurus to make the different terms for environment objects comparable between the information systems. A more detailed description of the usage of thesauri in Environment Information Systems can be found in (Kramer/Nikolai 1998).

The majority of internet-based Environment Information Systems use the WWW-UDK by the federal states of Germany and Austria. This ensures a more consistent usage of terms and concepts. However, the information systems store very different parts of environment data, often also in a varying level of detail. So the user has to navigate through keyword catalogues in order to find information on a certain topic. To navigate means, he has to enter a term or a combination of terms as query and assess the relevance of the system's answer. Then he can conclude how to refine his query in order to receive a more appropriate answer. This navigation is only possible with knowledge about semantic and contextual relations in the domain. The automated retrieval often fails or performs rather poorly because the search engines and information agents lack this understanding of context and semantics.

1.2 Semantic Navigation Maps

To overcome these problems, we have developed the concept of Semantic Navigation Maps. Semantic Navigation Maps can represent the complex logical relations within a knowledge base. This representation makes implicit semantic relations visible to the user and accessible for information agents. Thus, the agents can improve their search for information with knowledge about semantic relations in the domain.

The idea for this mechanism is based on the representation of ontological system semantics with self-organizing feature maps (SOM) by T. Kohonen (Kohonen 1997). A SOM is an artificial neural network which transforms feature vectors of objects, which form the input space of a Kohonen map, into a two-dimensional area of neurons that make up the SOM's output space, and preserves the logical neighborhood relations of the input space. The neighborhood relations in the map measure the semantic similarity of the objects. Thus, related context in different information sources can be found and categorized using a Semantic Navigation Map for the specific knowledge domain.

The remainder of this paper is organized as follows: in the next section we give a more detailed view of the automated information retrieval and present an example.

We model the knowledge space for this example, which is used for the training of our Semantic Navigation Map. In the third section we explain the foundations of our work, the self-organizing maps and information agents. Following we give an overview of related work. In section five we describe the creation of Semantic Navigation Maps and the application in information retrieval. Finally we discuss several problems of maintaining semantic distribution maps and possible extensions of the model.

2 Semantic relations in knowledge spaces

2.1 Semantic access to information

The motivation for our work is the problem of retrieving data regarding a limited context from several Environment Information Systems. To retrieve a data object directly by its designator, either there has to be a central replication of all data or a global schema which integrates all local schemata of the information systems must exist. The complete replication of all data however, is not feasible in most cases.

The integration of local schemata in a system-wide global schema was topic of numerous research projects in the domain of federated systems, especially federated databases. However, the loosely coupling of systems on the schema level massively increases the administrative overhead and consistency problems. Even the semantic and structural integration of local schemata is a very complex task which is not completely solved yet (Conrad 1997).

Therefore, in the following we consider only independent information systems whose local schemata are unknown to the user. For this reason most Environment Information Systems provide keyword catalogues and thesauri to support the information retrieval. Keywords are used as semantic access to the pure data objects. But because of the lack of a standardized keyword system, a single keyword is often not sufficient to designate a data object unambiguous. Even within the catalogues there exist synonyms to broaden the access to the data. Between different information systems problems with homonyms can arise and not all keywords may exist in every catalogue. Thus to designate a data object semantically unambiguous, not only the keyword is needed, but also some surrounding context. Several Environment Information Systems attempt to provide this context by establishing keyword hierarchies or semantic networks. With these tools the user can select the appropriate context for his query. However, automated information agents often do not have the required domain knowledge to navigate in keyword hierarchies and semantic networks or they are unable to recognize semantic similarities between different hierarchies. But if the agents should retrieve data from heterogeneous information sources, it is imperative that they can categorize the context of an answer in their own knowledge space.

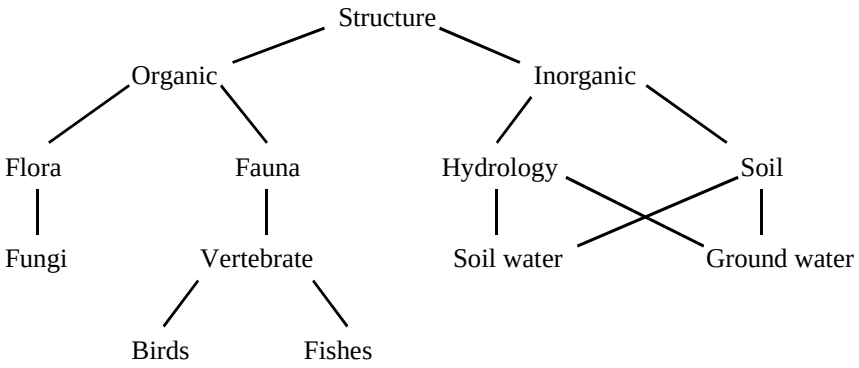

Figure 1
Part of KERIS' keyword hierarchy

2.2 Example of a knowledge space

To give an example of a knowledge space we use the Environment Information System KERIS (Heinrich/Hosenfeld 1997). The information system KERIS is public interface and integration point for environment data recorded by the Ecology Center of the Kiel University. KERIS provides a tool for geographical queries and an extensive keyword hierarchy. Entry points to this hierarchy are a set of general and abstract concepts. The hierarchy consists mostly of five levels of keywords with increasing specialization. Between the paths downward the hierarchy exist references. So the hierarchy resembles a directed graph rather than a tree. Figure 1 shows a small part of KERIS' keyword hierarchy.

The designators in the hierarchy are semantic keys to the data objects. The path from an entry point to a leaf determines the context of the referred data objects. Every sublevel limits the context more and therefore the keywords point to more special data objects. In this example we can assume to find in the Environment Information System data objects, which are structurally separable between organic and inorganic, flora and fauna and so on. For every path we can define the semantics by the keywords along the path. With this information we can generate for every path a feature vector in the form $V_j(p_1,p_2,...,p_n)$, where p_i states whether the path contains a certain context or not. The value of the p_i are therefore *TRUE* (1) or *FALSE* (0). Table 1 shows the seven 13-dimensional feature vectors of our example. Please note that the sequence of keywords in the feature vectors is independent from the sequence in the paths. The feature vector encode only the existence of keywords in a context, not

their order. Therefore it is possible to compare paths of different length and keyword order by means of feature vectors.

The feature vectors have certain similarities, which means, they partially describe a comparable semantic. In the directed graph this can be measured by the number of nodes belonging to two paths. The more nodes two paths share, the more do they describe a related context. In the example the path *Structure→Organic→Flora→...* describes structural information of organic entities, especially of the flora. The path *Structure→Organic→Fauna→...* describes comparable information of the fauna.

In the problem scope (or the knowledge space) spanned by the keywords, feature vector with similar semantics are spatially closer to each other than to such vectors with different semantics. Thus, an unknown data object described by a set of known keywords can be positioned in the knowledge space. From the spatial closeness to known context a similarity in the semantics of the data object can be concluded. This way, the representation of the knowledge space makes information of the semantics of data objects in the information system visible and exploitable.

To obtain a planar map of the neighborhood relations in the knowledge space we use Kohonen's self-organizing feature maps (SOM). The SOM architecture is an artificial neural network developed for the representation of semantic and functional relations of adjacent neurons in the human brain.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Structure	1	1	1	1	1	1	1
Organic	1	1	1	0	0	0	0
Inorganic	0	0	0	1	1	1	1
Flora	1	0	0	0	0	0	0
Fauna	0	1	1	0	0	0	0
Hydrology	0	0	0	1	1	0	0
Soil	0	0	0	0	0	1	1
Fungi	1	0	0	0	0	0	0
Vertebrate	0	1	1	0	0	0	0
Soil water	0	0	0	1	0	1	0
Ground water	0	0	0	0	1	0	1
Birds	0	1	0	0	0	0	0
Fishes	0	0	1	0	0	0	0

Table 1
Feature vectors of the keyword hierarchy

3 Foundations of this work

3.1 Kohonen's Self-Organizing Maps

The self-organizing neural network model was designed to map a multi-dimensional input space to a planar output space with preservation of the neighborhood relations of the input space. The output space however, is formed by a number of neurons (or units) usually arranged in a rectangular or hexagonal grid. It is quite possible to arrange the neurons in a hypercube and thus have a higher-dimensional output space with obviously better preservation of neighborhood relations. Anyway, most applications use the two-dimensional output for easy visualization and evaluation.

Each neuron has a number of synapse-weights which is equal to the dimension of the problem (or input) space. This specific weight-vector is the representation of the neuron in the input space while its grid coordinates define its place in the output space. The training of the SOM is an unsupervised learning method; i.e. there is no backpropagation of an error value. The learning algorithm itself lets the neurons align each other to represent the input space. Thus, the position of the neuron in the grid projects the topology of the input space onto the map. A more detailed description of the SOM and the learning algorithm can be found in (Kohonen 1997), (Ritter/Kohonen 1989), (Ritter et al. 1992).

Because the training of the SOM is unsupervised, a "good" representation of the input space requires representative training samples. Furthermore, a well-considered selection of map dimensions and training parameters is necessary. However, because the distribution of input data is usually not known, there exists no general definition for "good" representation by the map. Anyway, in unsupervised training this question does not arise because by default no target representation for quality assessment can exist. Also, no clear heuristics for the selection of training parameters do exist yet. Even the convergence of the algorithm is proven only for few cases (Cottrell et al. 1994), (Ruzicka/Hrycej 1993), (Tolat 1990).

To overcome the dilemma of fixed grid dimensions several architectures of growing networks are proposed (Fritzke 1991, 1995), (Rahmel 1996), (Burzevski/Mohan 1996). Unfortunately, these architectures are rather complex and perform not well with very high dimensional input data.

3.2 Information Agents

The concept of information agents, introduced in section 2.1 was developed in the research field of Distributed Artificial Intelligence. Information agents base on general mechanisms of software agents. In Artificial Intelligence software agents are viewed as modules for the autonomous solution of problems in distributed systems. However, there exists no standardized definition for software agents yet.

Klusch (Klusch 1998) summarizes the primary characteristics of software agents. For the purpose of our work we assume that agents have a certain grade of intelligence to store factual knowledge in a domain. Furthermore, they need the ability to process facts and reason. The agents should act autonomous and rational. Additionally they have to be mobile according to the requirements of distributed Environment Information Systems. Klusch describes information agents as agents which search actively and autonomous for relevant information in distributed information sources. Our work concentrates on the assessment of relevance.

4 Related work

4.1 Information retrieval in heterogeneous Environment Information Systems

To support the retrieval of data in heterogeneous Environment Information Systems Koschel et al. (Koschel et al. 1997) propose a system federation on the base of the CORBA standard. This requires all involved information sources to have a CORBA interface or at least provide an IDL description of their schemata. The local schemata are integrated and form an IDL interface. The resulting consistent access method facilitates an automated retrieval. However, a semantic support, as we intend it, is not provided.

Spiliopoulou, Faulstich and Roettgers (Spiliopoulou et al. 1997) use for the retrieval in Environment Information Systems a method derived from the data warehouse technique. The core of their system is a centralized Repository which stores an abstract schema of all available data. The retrieval of documents also uses a form of context matching. To enable this context matching, the system stores descriptions and meta-data of all information sources. A categorization of the information however, takes place only on the Repository level and not on the semantic level.

4.2 Semantic text classification

The application of SOM in semantic text classification describe for instance Merkl (Merkl 1997, Merkl/Schweighofer 1997) and Lin et al. (Lin et al. 1991). Merkl creates a semantic classification in the field of public international law. He expands the basic SOM to a multilevel hierarchy. The resulting semantic classification is used only as a supporting framework for human interaction. Contrary, we concentrate on the support of information agents.

The semantic text classification, especially by using the word context in sentences, is described in (Ritter/Kohonen 1989). Honkela and co-workers (Honkela et al. 1995) continued the experiment with words from Grimm tales. Later, they. (Honkela et al. 1996) applied a similar method for the semantic categorization of Usenet ar-

ticles. In our concept we use the technique to encode the semantic of data by the surrounding context. The resulting semantic maps are provided to information agents to process the represented semantic relations.

5 Support of the retrieval

5.1 Training of the Semantic Navigation Map

For our concept of Semantic Navigation Maps we use a re-implementation of the original SOM model described in section 3.1, merely with a slight modification to improve the representation abilities. Usually an activated neuron influences its neighborhood in all directions towards the input vector. From this point of view neurons in the corners and edges of the map have only a limited neighborhood. They have fewer neighbors than neurons in the center of the net. This leads to defects in the topology preservation of the mapping in these areas and apparently it appears an attraction of the edges. We often find the best matches to the training set along there and at the corners of the grid. To avoid this effect we view the rows and columns of the rectangular grid as circles – which is a simplification of projecting the map onto the surface of a sphere. Now each neuron always owns a complete neighborhood: Two distances vertically and two distances horizontally between each two neurons of the grid. From these distances we use the smaller one for neighborhood adaptations. This results in a better preservation of neighborhood relations of the input space as before. However, we have to regard that the left and right border of the map are adjacent as well as the upper and lower border and the four corners. In (Andreu et al. 1997) a similar model of toroidal SOM's is proposed to reduce errors in the topology preservation.

For the example in section 2 and the feature vectors in table 1 we can generate the following semantic map (figure 2).

We can see that similar knowledge domain are represented adjacently in the map, e.g. *Structure→Organic→Fauna→Vertebrate→Birds* (P2) and *Structure→ Organic→Fauna→Vertebrate→Fishes* (P3). Likely, the domains P4, P5 and P6 are represented relative closely. The domain *Structure→Organic→Flora→Fungi* (P1) is disjoined from all others and therefore its representatives are relatively isolated.

Where does the adjacent representation of similar input vectors in the output space come from? With the adaptation of it's neighborhood neuron *c* equalizes the weight vectors of the surrounding neurons to it's own weights in the training phase. The neighbors of *c* now represent similar vectors with their weights. Therefore an input vector which is similar to that represented in *c*, i.e. both are neighboring in the input space, will likely find it's best match in one of *c*'s neighbors.

The method of coding keywords we use here causes a severe problem with a very large number of keywords. With the described coding technique, a feature vector has

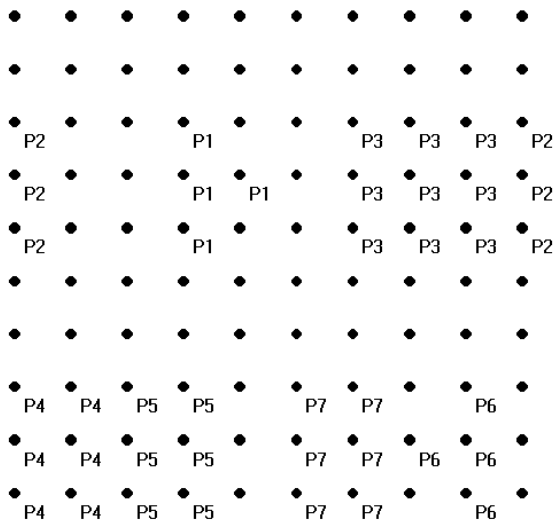

Figure 2
Representation of the keyword hierarchy by a SOM

a dimension equal to the number of keywords. The KERIS system provides more than 7000 keywords in its catalogue. With this amount of keywords a more sophisticated method of coding is required. Ritter and Kohonen propose a coding of single words. Every keyword is substituted by randomly generated vector. Then, a context is formed by the concatenation of the word vectors. Ritter and Kohonen (Ritter/Kohonen 1989) prove that these codes have the same stochastic features than the input space. Honkela et al. (Honkela et al. 1996) used this method later for a very large collection of keywords. In order to provide a better understanding of our approach we did intentionally not use this encoding technique in the KERIS example.

5.2 Semantic classification

The next step is to exploit the semantic map in order to categorize the context of given query results. From the spatial closeness to knowledge domains in the map the semantic similarity can be estimated. This similarity measure answers the question for the relevance of a query result.

As an example we give the query *Structure→Organic→Flora.* First, we position this query in our knowledge space (figure 3). The semantics of this query differs slightly from the context P1. In figure 3 the neuron with the smallest distance between the weight vector and the query vector is labeled with “Q”. The query vector

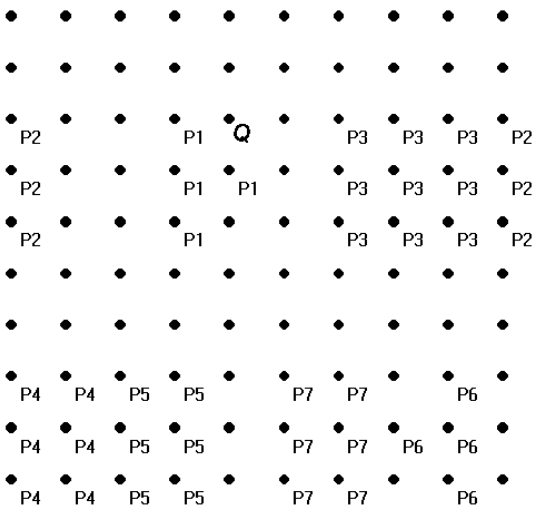

Figure 3
Position of the query in the semantic map

was created in the same way we generated the feature vectors for the training of the map.

Possible answers of different information systems might be: *Organic, Flora* (X1); *Organic, Flora, Fungi* (X2) ; *Flora, Fungi, Hydrology, Soil water* (X3). We generate a feature vector for every query result and give it as input to the semantic map.

The answer X1 (0,1,0,1,0,0,0,0,0,0,0,0,0) fits our query relatively good. Although the most general concept *Structure* is not contained, the same neuron representing our query is activated by the answer (figure 4).

The answer X2 (0,1,0,1,0,0,0,1,0,0,0,0,0) matches the concept P1 better than our query. Again the concept *Structure* is not contained. Because of the more limited context *Fungi* the answer differs slightly from the query context (figure 5).

The third answer X3 (0,0,0,1,0,1,0,1,0,1,0,0,0) differs notably from the query context (figure 6). Thus, especially the data objects of the first and second information system are relevant in the context of our query.

Using this relevance assessment the agent can make an intelligent selection of the most promising information source for a given query. The factual knowledge for the information agent, mentioned in section 3.2, can be provided with a Semantic Navigation Map. The ability to process factual knowledge comprises then the generation of feature vectors and evaluation of neighborhood relations in the map.

Obviously, we did not use the information provided by the hierarchical structure. Instead we only identify paths along the hierarchy and view them as knowledge do-

mains. While the usage of the hierarchy information might have improved the results in some places, we have to be aware of information sources, which are not able to organize their data in this way.

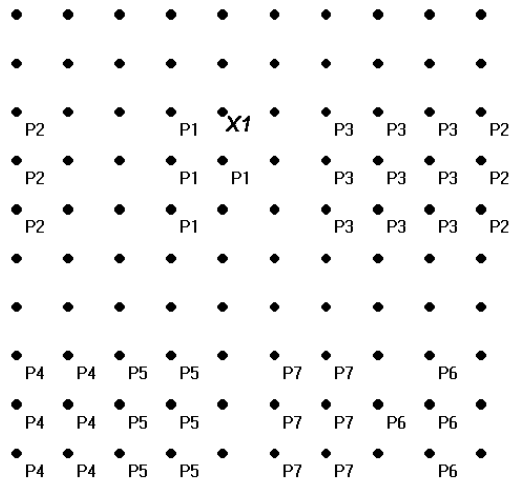

Figure 4
Position of the answer *Organic, Flora* in the map

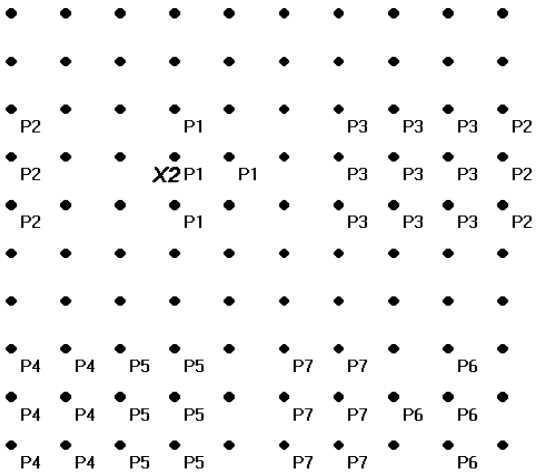

Figure 5
Position of the answer *Organic, Flora, Fungi* in the map

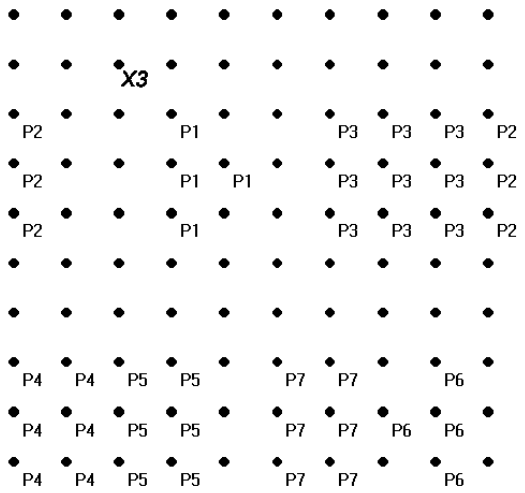

Figure 6
Position of the answer *Flora, Fungi, Hydrology, Soil water* in the map

6 Conclusion

In most cases the information retrieval in heterogeneous Environment Information Systems is based on the selection of data objects by their semantics. Therefore, Environment Information Systems provide extensive keyword catalogues to access the stored information. The meaning of keywords however, often depends on the surrounding context. Today, the automated intelligent retrieval of information is often hindered by semantically weak data models of information agents. The agents are not able to conceive semantics and context.

Our concept of Semantic Navigation Maps allows, as extension of regular system architectures, to make the semantics of a knowledge space visible and exploitable for information agents. On the base of self-organizing maps we can elicit the complex semantic relations within a knowledge space and represent them in a planar map. The semantic comparison between query results in the map answers the question which knowledge base most likely contains information related to the query. This allows to support the automated information retrieval with factual knowledge and semantic relations of the domain.

Several extensions of our approach are possible. It would be useful to process query results by a thesaurus to substitute synonymous keywords. Also, we have li-

mited the description of semantic similarities to the usage of keywords. If system specific meta-data, like models, relations or units of measurement, are available, their inclusion in the description of context could improve the classification of data objects. Currently we work on a hierarchical model of the Semantic Navigation Maps. We suppose a hierarchical map will develop a better representation of input data already grouped in a hierarchical form. Moreover, a hierarchical map should provide a better support for the navigation from general to more special concepts in the knowledge domain.

References

Andreu, G., Crespo, A., Valiente, J.M. (1997): Selecting the Toroidal Self-Organizing Feature Maps (TSOFM) Best Organized to Object Recognition, in: Proceedings of the ICNN-97 International Conference on Neural Networks, Vol. 2

Burzevski, V., Mohan, C.K. (1996): Hierarchical Growing Cell Structures, in: Proceedings of the IEEE International Conference on Neural Networks

Conrad, S. (1997): Federated Database Management Systems, Concepts of Data Integration (in German), Berlin et al.

Cottrell, M., Fort, J.C., Pagès, G. (1994): Two or three things that we know about the Kohonen algorithm, Technical Report, Université Paris 1, Number 31, Paris, France

Fritzke, B. (1991): Unsupervised clustering with growing cell structures, in: Proceedings of the IJCNN-91, Vol 2, Seattle

Fritzke, B. (1995): Growing Grid – a self-organizing network with constant neighborhood range and adaptation strength, in: Neural Processing Letters, 2, No. 5, S. 9-13

Heinrich, U., Hosenfeld, F. (1997): The Ecology Information System KERIS in the Internet (in German), in: Geiger, W. et al. (Eds.): Umweltinformatik '97, 11. International Symposium of the German Informatics Society (GI), Straßburg

Honkela, T., Kaski, S., Lagus, K., Kohonen, T. (1996): Newsgroup Exploration with WEBSOM Method and Browsing Interface, Helsinki Univ. of Technology, Faculty of Information Technology, Report A32

Honkela, T., Pulkki, V., Kohonen, T. (1995): Contextual Relations of Words in Grimm Tales, Analyzed by Self-Organizing Map, in: Fogelman-Soulie, F., Gallinari, P. (Eds.): Proc. of the Int'l. Conf. on Artificial Neural Networks, ICANN-95, EC2 et Cie, S. 3-7

Klusch, M. (1998): Cooperative Information Agents in the Internet (in German), Hamburg

Kohonen, T. (1997): Self-organizing maps, Berlin et al.

Koschel, A., Kramer, R., Schöckle, M., Schmidt, F., Spandl, H. (1997): Federation of heterogeneous Information Sources for Environment Information Systems (in German), in: Geiger, W. et al. (Eds.): Umweltinformatik '97, 11. International Symposium of the German Informatics Society (GI), Straßburg

Kramer, R., Nikolai, R. (1998): Technical and semantic Aspects of the loosely Integration of heterogeneous and autonomous Thesauri (in German), in: Proceedings of the German Informatics Society 1998 Workshop on heterogeneous active Environment Datebases

Lin, X., Soergel, D., Marchionini, G. (1991): A self-organizing semantic map for information retrieval, in: Proc. of the Int'l. Conf. on Research and Development in Information Retrieval, Chicago/IL

Merkl, D. (1997): Exploration of Text Collections with Hierarchical Feature Maps, in: 20th Annual Int'l. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'97), Philadelphia

Merkl, D., Schweighofer, E. (1997): The Exploration of Legal Text Corpora with Hierarchical Neural Networks: A Guided Tour in Public International Law, in: Proc. of the Int'l. Conf. of Artificial Intelligence and Law (ICAIL'97), Melbourne

Rahmel, J. (1996): SplitNet: Learning of Tree Structured Kohonen Chains, in: Proceedings of the ICNN-96, Washington

Ritter, H., Kohonen, T. (1989): Self-Organizing Semantic Maps, in: Biological Cybernetics, 61, No. 4, S. 241-254

Ritter, H., Martinez, Th., Schulten, K. (1992): Neural Computation and Self-Organizing Maps: An Introduction, Reading/MA

Ruzicka, P., Hrycej, D. (1993): Topological maps for invariant features repersentation and analysis of their self-organization, Neuro Nimes EC2, Nanterre, France

Spiliopoulou, M., Faulstich, L., Röttgers, J. (1997): A Concept of integrative Retrieval in independent Environment Information Systems (in German), in: Geiger, W. et al. (Eds.): Umweltinformatik '97, 11. International Symposium of the German Informatics Society (GI), Straßburg

Tolat, V.V. (1990): An analysis of Kohonen's self-organizing maps using a system of energy functions, in: Biological Cybernetics, 64, S. 155-164