Organization of Distributed Digital Libraries: A Neural Network – Based Approach

Andreas Rauber and Dieter Merkl

Department of Software Technology, Vienna University of Technology
Resselgasse 3 / 188, A-1040 Vienna, Austria.
e-mail: {rauber, merkl}@ifs.tuwien.ac.at

Abstract. Self-organizing maps are a popular neural network model for mapping high-dimensional input data onto a lower-dimensional output space. However, as the size of the training data increases, both the necessary computational power as well as the training time required exceed tolerable limits. Still more important, not all training data may be available in one central location but may rather be collected and managed at different sites. This paper describes an approach for combining independent, distributed self-organizing maps to build a higher order map, allowing the creation and maintenance of scalable, independent map systems, which can be built to suit the individual needs of the users. This is achieved by training higher order maps using the trained lower order maps as input data. We demonstrate the applicability of this approach in the field of digital libraries.

1 Introduction

Digital libraries, like their conventional counterparts, resemble, by their very nature, distributed sources of information. These individual libraries are managed and maintained independently and the organization of every library is suited to its specific needs. As the size and equally the speed of growth of these libraries increases, new tools are needed to organize these amounts of information. Among other approaches, neural networks have proven to be useful for coping with the kind of demands encountered in that application arena, namely adapting to a changing environment and dealing with the 'noise' of large amounts of unstructured information. One of the central goals of the application of neural networks in this domain is the content-oriented structuring of the single pieces of information stored in a library, which in turn is a starting point for query processing and library browsing interfaces. One special type of networks, namely the self-organizing map (SOM) [2] has been applied successfully in this domain on a number of occasions, be it the presentation of Usenet Newsgroup articles as in the *Websom* approach [1, 4], or the structuring and visualization of software libraries [5, 6]. However, all of these systems face the severe limitation, that all data must be available and processed at one place for producing a SOM representation of the library. This poses a severe restriction considering the natural structure of libraries as being distributed and independently managed. Apart from the requirement of having to combine these distributed sources of information and to integrate them into one single system, we are faced with the problem of scalability. The training of huge map systems requires by far too much processing time in order to allow frequent retraining, which becomes increasingly necessary as the speed of both growth and change of topics increases.

In this paper we describe an approach for a digital library representation based on self-organizing maps. The reason for using a SOM-based system for representing a collection of text documents is marked by its capability to preserve the topology of the input data, i.e. in the final SOM representation, similar documents are supposed to be located close to each other on the map. We suggest to use a set of several independently managed self-organizing maps, each of which represents a part of a network of digital libraries, which are integrated to form an overall view of the distributed library system. In a nutshell, this is achieved by using the trained SOMs as input to create a higher level SOM representing a collection of single libraries. However, the single maps need not necessarily be arranged in a strict hierarchical order, but can mutually reference each other, forming a Web-like structure, with the extent of this structure not being limited to a single location. Rather, distributed maps can be integrated and referenced via a network of systems, while all library maps can be maintained locally and independent from other referencing SOMs, allowing the retraining of the maps as the necessity arises. Furthermore, the approach is not limited to the use of the very standard SOM architecture for representing such a library, since most approaches based on a similar architecture can be integrated as well. This allows the user to choose the network architecture, network size and training frequency suiting his very needs.

The rest of the paper is organized as follows: Section 2 describes the basic architecture and training principles of the integrating SOM system. We then describe an application scenario in the field of text document classification in Section 3 and, finally, present our conclusions in Section 4.

2 The Architecture

The self-organizing map [3] consists of a grid of nodes with n-dimensional weight vectors, to which input signals $x \in \Re^n$ are presented in random order for training. An activation function based on some metric is used to determine the winning node (the 'winner') as the node showing the highest activation (e.g. the lowest Euclidean distance between the weight vector and the presented input signal). In the next step the weight vectors of the winner and its neighboring nodes are modified following some learning rate in order to represent the various input signals more closely, i.e. moving the weight vectors of the winner and of some neighboring nodes closer to the input pattern in order to increase their similarity in terms of the metric used. This finally forms a topology-preserving mapping of the high-dimensional input space.

Given a situation, where not all training data is available at one single location or where the amount of data hinders effective and frequent training, a series of small maps may be created based on the locally available subset of the complete data. Following the standard training procedure for the creation of these maps, they are integrated into a single higher order SOM.

To build a higher-order SOM integrating several lower order maps, we use the weight vectors of the lower order maps' nodes as input for a higher order SOM, resulting in a topology-preserving mapping of the nodes of the lower order maps. Thus we obtain a single map integrating the distributed maps.

Combining several independent SOM systems offers several advantages as compared to using one single, huge map. First of all, training times and computational requirements for training small maps, are considerably lower, making them more flexible. Secondly, small maps may easily and independently be retrained if the input data organized on the maps changes, which is especially important for dynamically changing environments, where new input data needs to be added to a trained map, e.g. when documents covering new subject matters are added to the digital library. Frequent retraining allows the mapping performance of the SOM to remain at high levels. Note, however, that a retraining of single lower order SOMs due to changes in the input data domain does not necessarily require a retraining of higher order maps where the 'new' domain may already be present, originating from the presence of this very domain in another lower order SOM. In this case the retrained lower order SOM only needs to be mapped onto the higher order map to determine the new node references. Thirdly, integrating several distributed maps allows the combination and utilization of SOMs, that are trained and maintained at different locations by downloading the weight vectors of the respective SOMs and using them to create a higher order map. Thus we can build systems on a higher level of abstraction based on locally maintained maps. Finally, the usage and integration of several smaller sets of maps provides a better comprehension of the underlying structure of large amounts of data whereas huge maps tend to become confused.

3 Application Scenario

As a digital library testbed we use the 1990 edition of the *CIA Worldfactbook*[1] which represents a text collection describing 245 countries, regions and islands of the world by their geographical, economical and political characteristics. To be used as input for training a SOM, each document is described by a feature vector based on a word histogram of the complete text corpus that is weighted following a simple $tf \times idf$ weighting scheme [7], i.e. term frequency times inverse document frequency. The indexing process identified 1056 unique terms excluding terms appearing in less than 15 or more than 220 documents, since they do not contribute to contents differentiation, leaving us with an input data set consisting of 245 feature vectors with 1056 elements.

For creating a number of first level SOMs the whole set of 245 documents was randomly split into 5 parts with each set comprising 50 documents, i.e. 5

[1] http://www.odci.gov/cia/publications/pubs.html

Burkina-Faso	Benin	Angola	Arctic Ocean	Atlantic Ocean		Bassas da India
Cameroon Cent. Afr. Rep. Chad	Burundi	Cap Verde		Antarctic	Bouvet Isl.	Clippert
Botswana	Bhutan	Cambodia		Andorra	Ashmore	Baker Isl.
Bangladesh	Burma	China	Albania		American Samoa	Christmas Isl. Cocos Isl.
Bahrain	Afghanistan	Algeria	Bulgaria	Brunei	Aruba	Anguilla
Colombia	Bolivia	Austria	Australia	Barbados	Belize	British Indian Ocean Territory
Brazil Chile	Argentina	Belgium	Canada	Antigua Bahamas	Bermuda	Cayman Isl.

Sweden	United Kingdom	United States	Iceland	St. Vincent Isl.	St. Lucia	Comoros
Austria Switzerland Japan	Thailad	Trinidad	Venezuela	St. Kitts & Nevis	Vanuatu	Tonga Tuvalu
Yugoslavia	Turkey	Sri Lanka	Suriname	Virgin Isl.	Turks & Caicos Isl.	Western Samoa
Vietnam	United Arab Emirates	Syria	Uruguay	St. Helens	Tokelau	St. Pierre & Miquelon
Taiwan	Sudan	Tunesia	Togo	Vatican		Wallis & Futuna
Uganda	Zaire	Tanzania	West Bank	Svalbard	Wake Isl.	
Zambia Zimbabwe	Swaziland	Mozambique	WORLD		Tromelin	Spratly Isl

Fig. 1. First Level Maps: 2 first-level maps out of a set of 5 SOMs consisting of 7×7 nodes each

documents are represented twice in different sets. Next, we independently trained 5 maps consisting of 7×7 nodes using these testsets. Due to space limitations we can only present two examples out of the complete set of 5 first-level maps in Figure 1. The remaining 3 maps, however, look quite similar in terms of their overall organization. Each field in the maps represents a node labeled with the names of the countries for which it is the best-matching representative, i.e. the winner. Units that were not winner for any country appear as empty fields.

Each of these 5 maps represents in itself a topologically ordered mapping of the corresponding documents, which means that countries considered similar to each other in terms of the facts given in the country description of the *CIA Worldfactbook*, are located on the same node or close to each other. In the lower left corner of the first map, for example, we find a number of nodes representing *south american countries*, which are followed by an *european* and *developed countries* area to the right. A cluster of *asian* and *african countries* is situated above the *south american* cluster. Another interesting cluster in the upper middle of the first map is represented by the *arctic* and *antarctic oceans*, followed by the *antarctic continent* and a number of *islands*. Another *european* cluster can be found in the upper left part of the map on the right hand side of Figure 1. Similar clusters may be found in all the other maps, e.g. clusters of *eastern european countries, countries of the arabic hemisphere*. Note, however, that the clustering provided by the mapping does not necessarily represent a geographical structuring of regions. Rather, the countries are organized on the map according to their overall similarity based on the descriptions in the *CIA Worldfactbook*.

In a second step, these 5 maps are integrated into one single SOM consisting of 7×7 nodes to represent the whole document collection. Thus we obtain a mapping of all nodes of the 5 lower level maps onto the nodes of the higher level SOM. The idea behind this approach is based on the fact, that nodes in

Kenya	Angola, Benin, Burkina Faso, Cameroon, C. Verde, CAR, Chad, Congo, Equ. Gu., Gabon, Gambia, Guinea, Guin. Bis., Ivory C., Madagasc, Malawi, Mali, Mauritania, Namibia, Niger, Sao T., Senegal, Sierra L, Tanzania, Togo, Uganda, Zaire, Zambia, Zimbabwe	Haiti	Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican R., Ecuador, El Salvador, Ethiopia, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, Venezuela	Fiji	United States	Australia, Austria 2x, Belgium, Canada, Denmark, Finland, France, Fed. Rep. of Germany, Greece, Italy, Japan 2x, Luxembourg, New Zealand, Norway, Portugal, South Korea, Spain, Sweden, Switzerland, United Kingdom
Ghana	Burundi, Rwanda	Mozambique 2x	Guyana, Suriname	Iceland 2x, Ireland	Bhutan, Nepal	Bulgaria, Czechoslovakia, German Dem Rep., Hungary, Poland, Romania, Soviet Union, Yugoslavia
Nigeria	Botswana, Lesotho, Swaziland	Seychell	Mongolia, North Korea, Taiwan, Vietnam	Jordan	Algeria, Bahrain, Djibouti, Egypt, Iraq, Israel, Kuwait, Lebanon, Libya, Morocco, North Yemen, Oman, Qatar, Saudi Arabia, Somalia, South Yemen, Sudan, Syria, Tunisia, United Arab Emirates	Albania, Cambodia, Laos
Comoros, 2x	Antigua, Bahamas, Barbados, Belize, Grenada, Jamaica, Mauritius, St. Kitts & Nevis, St. Lucia, St. Vincent & Grenandines, Trinidad	Papua New Guinea, Solomon Isl., Tonga, Tuvalu, Western Samoa	Andorra	Afghanis, Banglade, Brunei, Burma, China, Cyprus, India, Indonesia, Iran, Liberia, Malaysia, Maldives, Malta, Pakistan, Philippines, Singapore, Sri Lanka, Thailand, Turkey	South Africa	South Georgia
Macau	Vanuatu	American Samoa, Guam, Marshall, Micronesia, Northern Marian Isl., Palau, Puerto Rico, Virgin Isl.	Hong Kong, Netherlands	Faroe Isl., Gibraltar, Greenland	Arctic Ocean, Atlantic Ocean, Indian Ocean, Pacific Ocean	Spratly, Wake Isl.
Kiribati Nauru	Liechtenstein, Monaco	Antarcti	Gazastrip, West Bank	Ashmore, Coral Sea Isl., French Antilles	Iraq-Saudi Arabia Neutral Zone, Paracel Isl.	Johnston, Midway Isl.
Anguilla, Aruba, Bermuda, British Indian Ocean Terr., Cayman Isl., Christma, Cocos Isl., Cook Isl., Falkland, Guernsey, Isle of Man, Jersey, Montserrat, Niue, Norfolk Isl., Pitcairn, St. Helen, Tokelau, Turks & Caicos Isl.	San Marino, Vatican	French Guiana, French Polynesia, Guadeloupe, Martiniqe, Mayotte, New Caledonia, Reunion, St Pierre & Miquelon, Wallis & Futuna	Svalbard, WORLD	Jan Mayen	Bassas da India, Bouvet Isl., Clippert, Europa Isl., Glorioso, Heardmc, Juan de Nova Isl., Kingman Reef, Navassa Isl., Palmyra Atoll, Tromelin Isl.	Baker Isl., Howland Isl., Jarvis Isl.

Fig. 2. Second Level Map: 7×7 SOM integrating the 5 SOMs of Figure 1

the various lower level SOMs representing similar documents (e.g. the nodes representing oceans which are distributed across 3 lower level maps in our experimental setup) should be mapped onto one node in the higher level SOM, i.e. we should expect one cluster for every region described in the document collection. The corresponding higher level map is given in Figure 2. Note, that the main clusters are clearly visible from the map representation due to the accumulation of country descriptions on the cluster center nodes, with the nodes of the lower-level maps being mapped according to a higher level of abstraction. Thus we find a cluster of *african countries* in the upper left part of the map, with its center on the second node of the first row, being followed by a node representing *south american countries* in the middle of the first row. In the upper right corner of the map we find a node representing *western european* and *developed countries*, followed again by a node representing the former *communist hemisphere* below. Mind, that the documents used in these examples were taken from the 1990 edition of the *CIA Worldfactbook*, prior to the 'fall' of the communist hemisphere.

We further find several clusters of *islands* as well as a single node representing the *oceans* mentioned before, situated in the lower right part of the map. Please note that the countries that were present in the testbed twice (Austria, Comoros, Iceland, Japan and Mozambique) are now all mapped onto identical nodes. Generally, the nodes in the various maps representing highly similar or identical information, are now mapped onto the same area in the higher level representation of the text collection. Thus we find the higher level map to form an orderly mapping of all the input data used for training the single lower-order maps, with these maps now being integrated at a higher level of abstraction.

4 Conclusions

We have presented a novel approach for the integration of distributed self-organizing maps into higher-order maps by using the weight vectors of trained SOMs as input data to create a referencing map. Combining several independently managed maps offers itself to cope with situations where a number of distributed sources of information need to be managed independently, like it is the case in the digital library arena. Multiple small maps offer the additional advantage of being easier to retrain, thus providing greater flexibility, which is especially important in dynamically changing environments. Integrating several small SOMs allows the creation of large systems while keeping the overhead and maintenance cost low due to the distributed structure. The referencing SOMs need not necessarily form a strict hierarchy. Rather, selected maps of various orders can be integrated into any single map, allowing users to build their personal library systems including maps or even portions of maps that are relevant to their needs. Since the principle of combining these distributed sources are based on a rather generic vector-oriented representation, different types of systems may be integrated, providing a high degree of flexibility.

References

1. S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. Creating an order in digital libraries with self-organizing maps. In *Proc. of the World Congr. on Neural Networks (WCNN96)*, San Diego, USA, 1996.
2. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982. Springer Verlag, Berlin, Heidelberg, New York.
3. T. Kohonen. *Self-Organizing Maps.* Springer Verlag, Berlin, Germany, 1995.
4. T. Kohonen, S. Kaski, K. Lagus, and T. Honkela. Very large two-level SOM for the browsing of newsgroups. In *Proceedings of ICANN96, International Conference on Artificial Neural Networks, Bochum, Germany, July 16-19, 1996*, Lecture Notes in Computer Science, vol. 1112, pages 269–274. Springer, Berlin, 1996.
5. D. Merkl. A connectionist view on document classification. In *Proceedings of the 6th Australasian Database Conf. (ADC'95)*, Adelaide, SA, 1995.
6. A. Rauber and D. Merkl. Finding structure in text archives. In *Proc. European Symp. on Artificial Neural Networks (ESANN98)*, Bruges, Belgium, 1998.
7. G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA, 1989.

A. Rauber and D. Merkl.
Organization of Distributed Digital Libraries: A Neural Network – Based Approach
In *Proceedings of the 1st International Symposium on Intelligent Data Engineering and Learning*, 14. - 16. Oct. 1998, Hong Kong
L. Xu, L. W. Chan, I. King, a. Fu (Eds.) Springer–Verlag, 1998.