Creating an Order in Distributed Digital Libraries by Integrating Independent Self-Organizing Maps

Andreas Rauber, Dieter Merkl
Institut für Softwaretechnik, Technische Universität Wien
Resselgasse 3/188, A–1040 Wien, Austria
{andi, dieter}@ifs.tuwien.ac.at

Abstract

Digital document libraries are an almost perfect application arena for unsupervised neural networks. This because many of the operations computers have to perform on text documents are classification tasks based on "noisy" input patterns. The "noise" arises because of the known inaccuracy of mapping natural language to an indexing vocabulary representing the contents of the documents. A growing number of papers is dedicated to the usage of self-organizing maps to organize the contents of such digital libraries. These papers assume the central availability of the data; an assumption that is questionable given the massive amount of available information. In this paper we describe an approach for organizing distributed digital libraries based on a system of independent self-organizing maps each of which representing just a portion of the complete digital library. Furthermore, we argue in favor of integrating these independent maps in a hierarchical fashion, again by means of self-organizing maps. The integration is based on the trained low level maps.

1 Introduction

During recent years we have witnessed the appearance of an ever increasing flood of miscellaneous written information available in computer accessible form and originating from very different sources, culminating in the advent of massive digital libraries. As an example of such a digital library consider *The Internet Public Library* (http://www.ipl.org). Powerful methods for organizing, exploring, and searching collections of text documents are thus needed.

The map metaphor for displaying the contents of a document library in a two-dimensional display has gained considerable interest [1, 4, 5, 6, 7]. Maps are used to visualize the similarity between documents in terms of distance within the two-dimensional map display. Hence, similar documents may be found in neighboring parts of the map display.

The underlying assumption in all of the above mentioned papers is that the data items, i.e. the documents, are available locally for self-organizing map training. With ever growing information repositories, however, this assumption is questionable. One should rather expect that several repositories exist at different sites and a way for unified access should be provided for the users of such distributed digital libraries. Moreover, most of the reports on the application of self-organizing maps for the organization of document libraries propose the usage of a single map for document space visualization. It is only natural that with increasing size of the document library the self-organizing maps for representing this library grow larger, see e.g. [4]. This, however, does not necessarily improve the organization of the library. In particular, it is rather difficult for the user to find her orientation in large maps. To use an example from geography, it is much easier to find the way, say, from *Stockholm* to *Skövde* by using the map of *Sweden* than by using the map of *Europe* or, even worse, that of the *World*. In much the same way it is easier to locate a particular document using a map of a smaller portion of the library while relying on higher aggregated maps for overall orientation.

In this paper we suggest the utilization of individual self-organizing maps for organizing the different parts of a distributed digital library. Multiple maps may then be integrated to produce an aggregated visualization of the complete document space. We demonstrate the effects of such an approach based on a sample library of text documents.

The remainder of this paper is organized as follows. In Section 2 we give a brief review of self-organizing maps and describe their integration. Section 3 is dedicated to a description of our experimental document library. Experimental results of our approach to digital library organization are given in Section 4. Finally, our conclusions are contained in Section 5.

2 Hierarchies of self-organizing maps

The self-organizing map [2, 3], SOM, is one of the most prominent neural network models adhering to the unsupervised learning paradigm. The model consists of p units, each of which is assigned an n-dimensional weight vector m_i, $1 \leq i \leq p$. Training is performed as a repetition of (i) input pattern presentation, (ii) selection of the unit with the closest weight vector, i.e. the *winner*, and (iii) adaptation of the weight vectors of the *winner* and of a number of units in the neighborhood of the *winner*. It is important to guarantee that both the strength of adaptation as well as the number of units that are adapted apart from the *winner* are decreasing in time. The distinctive feature of the self-organizing map is that similar input patterns will be arranged in neighboring regions of the resulting map. Hence, the similarity between input data items, i.e. the documents in our application, is mirrored in terms of the distance of the respective *winners* within the map.

In case of a digital library that exists only distributed over several sites, it might be more efficient to have independent self-organizing maps that represent the various parts of the digital library than transfering the whole information to one site for training. However, when some form of uniform access to the data is requested, the contents of the various sites has to be integrated. With our approach to digital library organization we suggest to utilize self-organizing maps to perform such an integration. In particular, the map that shall integrate different portions of the digital library may be trained by using the weight

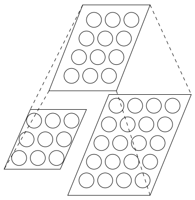

Figure 1: Integration of two self-organizing maps

vectors of the maps to be integrated. Such a strategy may be applied recursively in order to build hierarchies of arbitrary depth as shown in Figure 1. In this figure a 3×3 and a 4×5 map are integrated in a 3×4 map. Note that also selected parts of self-organizing maps may be integrated by using essentially the same architecture. The user simply selects areas of interest scattered across different maps for which an integration shall be performed. By this, the user may tie together pieces of information to build her own library fine-tuned to her particular interests.

The effect of such an integration, obviously, is that input data items that are separated in different low level maps are grouped together in the high level map. Input data that are mapped onto the same low level unit are represented together in the high level map.

3 An experimental document archive

For the experiments presented thereafter we used the 1990 edition of the *CIA Worldfactbook* (http://www.odci.gov/cia/publications/factbook/index.html) as a sample document archive. The *CIA Worldfactbook* represents a document collection containing information on countries and world regions. The information is split into categories such as *Geography*, *People*, *Economy*, *Defense Forces*, etc. In total, the 1990 edition of the *CIA Worldfactbook* consists of 245 documents.

The various documents are represented by means of simple histograms of word occurrences. We used all words occuring in more than 15 and less than 220 documents. Thus, 1056 distinct words, i.e. index terms, remained that were weighted according to a $tf \times idf$ weighting scheme [8], i.e. term frequency times inverse document frequency. Such a weighting scheme assigns high weights to index terms that occur frequently within a document but rarely within the whole document collection. Finally, the documents are represented by feature vectors where each feature corresponds to an index term and the specific value of the feature is derived by means of the $tf \times idf$ weighting scheme. These feature vectors are used as the input data during network training.

German Dem Rep	Finland	German Fed Rep	France	Iran	Howland Isl.	Europa Isl. Glorioso Isl.
Czechoslovakia Hungary	Denmark	Egypt	Greece		Coral Sea Isl.	Heard Isl.
Cuba	Ecuador	Honduras	Hong Kong	Cyprus		French Antilles
Costa Rica El Salvador Ethiopia Guatemala	Dominican Rep.	Haiti	Fiji	Falkland Isl.	Gibraltar	
Guyana	Guinea Bissau	Djibouti	Grenada	Cook Isl.	Faroe Isl.	Greenland
Congo	Equatorial Guinea	Gambia	Comoros	French Polynesia	Guam	Guernsey
Ghana	Gabon	Guinea	Guadeloupe	French Guyana		Gaza Strip

Figure 2: Low level self-organizing map

4 Experimental results

In order to simulate a distributed digital library, the documents of the *CIA Worldfactbook* have been split randomly into five parts consisting of 50 documents each. Note that five documents are thus assigned to two parts at a time. Each part of the library was then used for training of a separate 7×7 self-organizing map. Each of these maps represents a topologically ordered portion of its subset of the document library. Due to space restrictions we can only show one of these maps in this paper, see Figure 2. The other four maps, however, are rather similar in spirit.

In a second step, we integrated the five independent maps into one single 7×7 self-organizing map which now represents the complete document library. The map that results from this integration is shown in Figure 3. Note that the main clusters are clearly visible from the map representation, with the information from the low level maps being arranged according to the organizational principles of the high level map. As examples consider the area containing African countries in the left upper part of the map, the area of oil producing countries, the area of countries from Latin America, or the area of countries usually referred to as the *first world* in the right upper part. With the latter area, please note the explicit distinction into countries belonging to the Western hemisphere and those belonging to the Eastern hemisphere[1]. Finally, it is worth noting that the countries that are contained twice in the library are pairwise assigned to the same unit in the integrated map. These countries are *Austria*, *Comoros*, *Iceland*, *Japan*, and *Mozambique*.

In a nutshell, the higher level map forms an orderly mapping based on the

[1] At this point it is important to recall that our document archive is the 1990 edition of the *CIA Worldfactbook*. Thus, the descriptions refer to a time before the "fall" of the Communist hemisphere.

Kenya	Angola, Benin, Burkina Faso, Cemeroon, Cape Verde, Centr Afric Rep, Chad, Congo, Equat Guinea, Gabon, Gambia, Guinea, Guinea Bissau, Ivory Coast, Madagascar, Malawi, Mali, Mauritania, Namibia, Niger, Sao Tome, Tanzania, Togo, Uganda, Zaire, Zambia, Zimbabwe	Haiti	Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Rep, Ecuador, El Salvador, Ethiopia, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, Venezuela	Fiji	United States	Australia, Austria (2x), Belgium, Canada, Denmark, Finland, France, German Fed Rep, 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	Seychelles	Mongolia, North Korea, Taiwan, Vietnam	Jordan	Algeria, Bahrain, Djibouti, Egypt, Iraq, Israel, Kuweit, Lebanon, Libya, Morocco, North Yemen, Oman, Qatar, Saudia Arabia, Somalia, South Yemen, Sudan, Syria, Tunisia, UAE	Albania, Cambodia, Laos
Comoros (2x)	Antigua, Bahamas, Barbados, Belize, Grenada, Jamaica, Mauritius, St Kitts, St Lucia, St Vincent, Trinidad	Papua New Guinea, Solomon Isl, Tonga, Tuvalu, Western Samoa	Andorra	Afghanistan, Bangladesh, Brunei, Burma, China, Cyprus, India, Indonesia, Iran, Liberia, Malaysia, Maldives, Malta, Pakistan, Philippines, Singapore, Sri Lanka, Thailand, Turkey	South Africa	South Georgia
Macau	Vanuatu	Americ Samoa, Guam, Marshall Isl, Micronesia, N Mariana Isl, Palau, Puerto Rico, Virgin Isl	Hong Kong, Netherlands	Faroe Isl, Gibraltar, Greenland	Arctic Ocean, Atlantic Ocean, Indian Ocean, Pacific Ocean	Spratly Isl, Wake Isl
Kiribati, Nauru	Liechtenstein, Monaco	Antarctica	Gaza Strip, West Bank	Ashmore Isl, Coral Sea Isl, French Antarctic Lands	iraq-SA-Neutr Zone, Paracel Isl	Johnston Atoll, Midway Isl
Anguilla, Aruba, Bermuda, British Indian Ocean Terr, Cayman Isl, Christmas Isl, Cocos Isl, Cook Isl, Falkland Isl, Guernsey, Isle of Man, Jersey, Montserrat, Niue, Norfolk Isl, Pitcairn Isl, St Helena, Tokelau, Turks Isl	San Marino, Vatican	French Guiana, French Polynesia, Guadeloupe, Martinique, Mayotte, New Caledonia, Reunion, St Pierre, Wallis & Futuna	Svalbard, World	Jan Mayen	Bassas da India, Bouvet Isl, Clipperton Isl, Europa Isl, Glorioso Isl, Heard Isl, Juan de Nova Isl, Kingman Reef, Navassa Isl, Palmyra Atoll, Tromelin Isl	Baker Isl, Howland Isl, Jarvis Isl

Figure 3: Integration of the low level self-organizing maps

information contained in several low level maps, each of which representing a portion of the library. The higher level map thus is a convenient starting point for a user trying to find her orientation in a distributed digital library.

5 Conclusions

In this paper we presented a novel approach for organizing distributed digital libraries by using a system of self-organizing maps. In our approach, different portions of a digital library that may reside on different sites are independently organized by self-organizing maps. These individual maps may be integrated in order to provide an overview picture of the information contained in the digital library. We suggested that the integration of low level maps should be based on the various weight vectors representing groups of documents. Such an integration, however, is not restricted to complete maps. The user may equally well define areas of interest that cover just portions of low level maps. These areas can then be integrated by using the very same integration process. The benefit for the user of such a distributed digital library is that she may define her own personal library that reflects her particular interests.

References

[1] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, Creating an order in digital libraries with self-organizing maps, In: Proceedings of the World Congress on Neural Networks, San Diego, CA, 1996.

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

[3] T. Kohonen, Self-Organizing Maps, Springer-Verlag, Berlin, 1995.

[4] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela, Very large two-level SOM for the browsing of newsgroups, In: Proceedings of the Int'l Conference on Artificial Neural Networks, Bochum, Germany, 1996.

[5] X. Lin, D. Soergel, and G. Marchionini, A self-organizing semantic map for information retrieval, In: Proceedings of the Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, IL, 1991.

[6] D. Merkl, A connectionist view of document classification, In: Proceedings of the Australasian Database Conference, Adelaide, SA, 1995.

[7] D. Merkl, Exploration of text collections with hierarchical feature maps, In: Proceedings of the Int'l ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, 1997.

[8] G. Salton, Automatic Text Processing: The transformation, analysis, and retrieval of information by computer, Addison-Wesley, Reading, MA, 1989.

A. Rauber and D. Merkl.
Creating an Order in Distributed Digital Libraries by Integrating Independent Self-Organizing Maps.
In *Proceedings of the 8th Int'l Conference on Artificial Neural Networks* (ICANN'98), Skövde, Sweden, September 2–4, 1998.
Springer–Verlag, Perspectives in Neural Computing.