SOMLib: A Digital Library System Based on Neural Networks

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

ABSTRACT

Digital Libraries have gained tremendous interest with numerous research projects addressing the wealth of challenges in this field. While computational intelligence systems are being used for specific tasks in this arena, the majority of projects relies on conventional techniques for the basic structure of the library itself. With the SOMLib project we create a digital library system that uses a neural network-based core for library representation and query processing. The self-organizing map, a popular unsupervised neural network model, is used to automatically structure a document collection. Based on this core, additional modules integrate distributed libraries and create an intuitive representation of the library, automatically labeling the various topical sections in the document collection.

Keywords

Self-Organizing Map (SOM), Document Clustering, Visualization, Distributed Digital Libraries

1 INTRODUCTION

The *SOMLib* digital library system is based on the self-organizing map, a popular unsupervised neural network architecture providing a topology-preserving mapping from a high-dimensional input space onto a two-dimensional output space. It is used to structure document collections on a 2-dimensional map locating documents on similar topics close to each other. Documents are transformed into a vector space representation and used for network training. To allow the creation of large library systems, instead of training huge single SOMs, individual maps can be integrated to form a set of referencing library maps. The various topical sections in the library are labeled automatically with the keywords assigned by the *LabelSOM* method. Last, but not least, the *libViewer* provides an intuitive user interface employing Dublin-Core based metaphor graphics. For the experiments presented in this paper we use the classic *TIME Magazine* article collection consisting of 420 articles from the *TIME Magazine* of the 1960's.[1]

2 SOM AND DIGITAL LIBRARIES

The self-organizing map (SOM) [2] is an unsupervised neural network model that provides a topology-preserving mapping from a high-dimensional input space to a usually 2-dimensional output space. It consists of a grid of units with n-dimensional weight vectors. During the training process input data are presented to the map in random order. An activation function based on some metric is used to determine the winning unit. Next, the weight vectors of the winner and of neighboring units are modified to represent the presented input signal more closely. The SOM has been applied successfully to text classification, cf. [1, 3]. Text documents can be thought of topical clusters in the high-dimensional feature space spanned by the individual words in the documents. Full-term indexing is used to represent the documents, using a $tf \times idf$, i.e. term frequency times inverse document frequency, weighting scheme [6]. A trained SOM thus represents a topological ordering of the documents, meaning that documents on similar topics are located close to each other on the 2-dimensional map. This is comparable to what one can expect from a conventional library, where we also find the various books ordered by some contents-based criteria.

A query presented to a *SOMLib* is mapped onto the corresponding map location, retrieving the documents mapped onto the respective unit and facilitating intuitive browsing of similar documents located on neighboring nodes due to the clustering capability of the SOM. Again, this is what commonly happens in a conventional library when the librarian points you to the section of your interest in the library building.

3 DISTRIBUTED COLLECTIONS

For training a SOM it is assumed that all input data is available locally to be presented to the map. This is not the case in the real world where document collections either exist as several independent libraries or are released at different points in

[1] The article collection as well as the experimental results described in this paper are available for interactive exploration at http://www.ifs.tuwien.ac.at/ifs/research/ir.

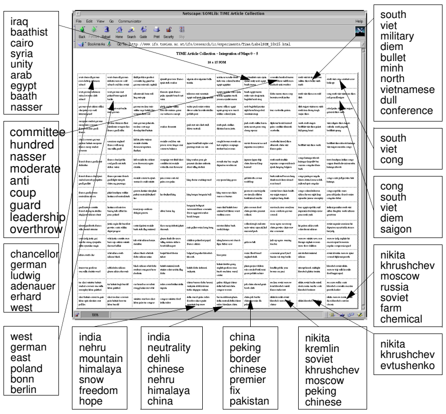

Figure 1: Labeled 10 $\times$ 15 SOM integrating 6 maps

time like periodicals. In both cases we want to integrate these distributed collections without having to face the overhead of training the whole SOM network again. Furthermore, as the number of topics to be represented by the SOM increases, the map also has to grow, leading to very big maps with all the problems like long training times etc. incurred. Both situations call for a way to integrate different, independent SOMs as opposed to other approaches relying on one central large SOM.

To allow the integration of distributed libraries, instead of using the very document representations as input to the SOM, we use the weight vectors of existing SOM-based libraries to train a higher-level *SOMLib* map, thus allowing the convenient and independent integration of distributed libraries [5]. Every user can choose to create a personal library based on her individual interests by integrating various maps or even just parts of maps. An example for such an integrating map is provided in Figure 1, integrating 6 independently trained maps. Each of these 6 maps represents between 53 and 87 articles from the *TIME Magazine*, with the integrating map representing the whole article collection of 420 articles.

4 LabelSOM: LABELING THE LIBRARY

While the SOM provides an intuitively ordered representation of the data, interpreting a trained SOM remains a somewhat tedious task in spite of the recent development of enhanced visualization techniques for SOMs. This is because the reasons for a specific cluster assignment are not explicit in the standard SOM representation, requiring the map to be labeled manually. The *LabelSOM* method [4] addresses this problem by automatically assigning labels to the units of the SOM describing the features of the data points mapped onto the respective unit. This results in a labeled SOM giving the contents of the documents in the various areas of the map, allowing the map to be actually read similar to a map of a conventional library. Figure 1 provides some labels automatically created for the *TIME Magazine* map, clearly identifying the topics of the various sections on the map.

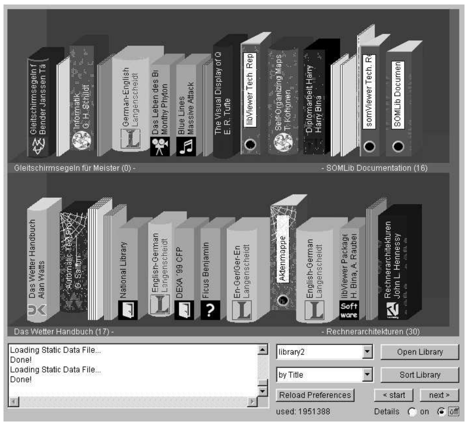

Figure 2: Visualizing metadata of documents

5 libViewer: VISUALIZING THE LIBRARY

The *libViewer* module provides an intuitive user interface to the *SOMLib* system by combining the spatial organization of documents provided by the SOM with a graphical interpretation of metadata based on the Dublin Core Metadata. Documents are assigned a physical representation template such as books, binders, papers etc., with further metadata such as language, date of last reference etc. being encoded by a set of additional metaphors. An example of a *libViewer* library representation is provided in Figure 2, using different physical representations for the various types of documents in the library, with additional information like language, amount of information, last time of reference etc. being visually encoded using different colors, thickness of documents, dust and spiderwebs etc.

REFERENCES

1. T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. WEBSOM - Self-organizing maps of document collections. In *Proc. Workshop on Self-Organizing Maps (WSOM97)*, Espoo, Finland, 1997.

2. T. Kohonen. *Self-Organizing Maps*. Springer Verlag, Berlin, Germany, 1995.

3. D. Merkl and A. Rauber. Uncovering associations between documents. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, Sweden, 1999.

4. A. Rauber. LabelSOM: On the labeling of self-organizing maps. In *Proc. International Joint Conference on Neural Networks*, Washington, DC, 1999.

5. A. Rauber and D. Merkl. Creating an order in distributed digital libraries by integrating independent self-organizing maps. In *Proc. Int'l Conf. on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.

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