Finding Structure in Text Archives

Andreas Rauber, Dieter Merkl

Institute of Software Technology
Vienna University of Technology
Resselgasse 3/188, A–1040 Vienna, Austria
{andi, dieter}@ifs.tuwien.ac.at

Abstract. With the advance and massive growth of electronic text archives, the need for tools emerges, which help to gain insight into the basic structure of the underlying digital library. We present a neural network approach for the analysis and exploration of text archives aiming at the detection and visualization of the inherent structure of the text collection. This cluster visualization technique called *Adaptive Coordinates* is based on an extended learning rule for the self-organizing map. It provides an intuitive visualization by mapping clusters in a high-dimensional input space onto groups of nodes in a 2-dimensional output space. We further compare the results of this mapping with another prominent cluster visualization technique, namely *Sammon's Mapping*.

1. Introduction

Traditional text archives exhibit a kind of structure, which allows the user to understand the overall organization of the text collection and provides a means to search and to browse the collection to retrieve relevant texts. However, with the increasing amount of electronically available text collections, exploration of those electronic text archives becomes a challenging task both for users of large text corpora as well as for researchers trying to provide the means necessary for intuitive analysis. Generally, exploration is not primarily a problem of query processing and retrieval of relevant documents, but rather one comprising the whole complex of understanding the text collection and its structure at a higher level of abstraction. Basically, the individual texts present in any text collection span a high-dimensional input space defined by the words occurring in the various text documents. The goal is to provide a method that allows easy and intuitive access as well as aids in the understanding of this high-dimensional document space, enabling both the retrieval of documents based on queries as well as interactive browsing to locate relevant documents and to make the overall structure of the text collection intelligible to the user.

Numerous approaches to the problem of structure analysis of text corpora have been developed, either trying to impose a hierarchy on a given text collection or to provide some other way of clustering, using both supervised or unsupervised analysis methods [6]. However, most systems primarily provide a method for convenient and 'intelligent' document retrieval based on query systems of differing degrees of sophistication with too little emphasis on visualization so far. As a consequence, interactive exploration is usually not supported. One well-known technique for the visualization of high-dimensional data spaces is *Sammon's Mapping* (SM) [7], aiming to represent the distances between data points in the high-dimensional input space as closely as possible in a 2-dimensional plot. Recent approaches use neural networks to structure large text corpora and to provide an interface for intuitive browsing of these collections. A prominent neural network architecture based on unsupervised learning is the self-organizing map (SOM), which has repeatedly been used to analyze and to visualize text archives, the most prominent example probably being the WEBSOM project [3].

The standard map display to represent the results of SOM training has its limits in that cluster boundaries are difficult to detect. To overcome this problem, we apply a new visualization technique based on an extended learning rule for SOM resulting in an intuitive representation of clusters as groups of nodes in a 2-dimensional output space. The basic idea of this *Adaptive Coordinate* (AC) approach [5] is to have the nodes of the SOM arrange themselves in a 2-dimensional output space during the training process in such a way as to approximate their geometric relationship in the high-dimensional vector space as faithfully as possible. The resulting visualization of the trained SOM is by its very idea similar to the SM, but stems from the self-organization during the learning process.

In this paper we demonstrate the application of SOM enhanced with the AC visualization technique to the problem of structure visualization of free form text corpora. We further compare the resulting AC visualization both with the standard SOM visualization as well as with the corresponding SM to analyze its capabilities in the fields of text archive exploration.

2. Sammon's Mapping

Sammon's Mapping (SM) [7] provides a mapping from a high-dimensional vector space onto a 2-dimensional output space. The basic idea is to arrange all the data points on a 2-dimensional plane in such a way, that the distances between the data points in this output plane resemble the distances in vector space as defined by some metric as faithfully as possible. More formally, given a set of data points x_i in $\Re^n$ with $d(x_i, x_j)$ being the distance between two data points according to some metric, we obtain a distance matrix D with elements d_{ij} in input space. Let o_i be the image of the data item x_i in the 2-dimensional output space. With O we denote the distance matrix containing the pairwise distances between images as measured by the Euclidean vector norm $\|o_i - o_j\|$. The goal is to place the o_i in such a way that the distance matrix O resembles as closely as possible matrix D, i.e. to optimize an error function E by following an interative steepest-descent process.

$$E = \frac{1}{\sum_i \sum_{j>i} d_{ij}} \sum_i \sum_{j>i} \frac{(d_{ij} - \|o_i - o_j\|)^2}{d_{ij}} \tag{1}$$

The resulting visualization depicts clusters in input space as groups of data points mapped close to each other in the output plane. Thus, the inherent structure of the input signals can be told from the structure detected in the 2-dimensional visualization.

3. SOM and Adaptive Coordinates

The *Adaptive Coordinates* (AC) approach is an extension to the standard learning procedure for Kohonen's self-organizing map (SOM) [2]. The standard training procedure itself remains unmodified as follows: Input signals $x \in \Re^n$ are presented to the map, consisting of a grid of units with n-dimensional weight vectors, in random order. An activation function based on some metric (e.g. the Euclidean Distance) is used to determine the winning unit (the 'winner'). In the next step the weight vector of the winner as well as the weight vectors of the neighboring units are modified following some learning rate in order to represent the presented input signal more closely.

The basic idea of the AC approach is to visualize clusters in input space in a 2-dimensional output space. By extending the basic training procedure we try to mirror the movement of the nodes' weight vectors in the high-dimensional input space in a 2-dimensional output space. For mirroring this movement, each unit i is assigned a position in this output space, with its position as given by two adaptive coordinates $\langle ax_i, ay_i \rangle$ initially being identical to the unit's position in the map grid. During the learning process, at each step t the distances between each unit and the presented input signal are stored in a table $Dist(t)$. After adapting the weight vectors following the standard training rule, this distance table is calculated again for the same input signal as $Dist(t+1)$. Based on these tables the relative change in distance between the weight vector of every unit i and the presented input signal can be calculated as

$$\Delta Dist_i(t+1) = \frac{Dist_i(t) - Dist_i(t+1)}{Dist_i(t)} \tag{2}$$

describing the movement of the unit's weight vectors towards the presented input signal in input space. This movement is now performed analogously in the 2-dimensional output space with the winning unit being representative of the presented input signal, i.e. the adaptive coordinates of every unit but the winner are modified such that the unit's new position is moved by the same fraction as given in $\Delta Dist$ towards the winner's position given by $\langle ax_c, ay_c \rangle$. Thus, the movement of the adaptive coordinate ax_i can be given as

$$ax_i(t+1) = ax_i(t) + \Delta Dist_i(t+1) \cdot (ax_c(t) - ax_i(t)) \tag{3}$$

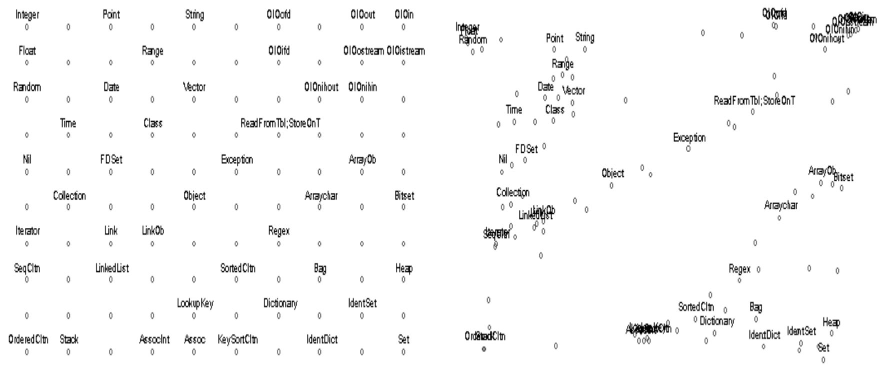

Figure 1: 10×10 SOM of the NIHCL

with the adaption of ay_i being performed analogously. Using the adaptive coordinates $< ax_i, ay_i >$ to plot the units location in the 2-dimensional output space allows the visualization of the clustering learned by the SOM.

4. NIHCL Text Archive Exploration

The following experimental results are based on the manual pages of the NIH C++ class library collection (NIHCL) [1] as a sample text archive. The library consists of a number of classes, ranging from input/output operations to general data types and container classes. Binary vector representations were created by full-text indexing the manual pages, where 1 indicates the presence of a specific word and 0 its absence in the manual page of a specific class, with 489 distinct words being detected in the manual page collection. This resulted in input vectors of dimensionality 489 being used as input signals to both the SM as well as the SOM training process.

The left part of Figure 1 depicts the standard representation of a 10×10 SOM trained with the NIHCL data. The basic structure of the NIHCL is present in the resulting mapping with, for example, all classes concerning file input/output operations like *OIOin, OIOout* being mapped onto the upper right corner of the map. Data types like *Integer, Point, String* are mapped onto the upper left part. However, we have to admit that the detection of these details of the structure and the extraction of further groupings of classes is hardly possible without profound knowledge about the classes themselves. Consider, for example, the classes *Point, String* and *OIOofd*. We have identified *String* to be a data type like *Point*, whereas *OIOofd* is a class dealing with input/output operations. However, *String* is separated from both *Point* and *OIOofd* by a single blank node. Thus, one might come up with the erroneous perception that all three classes are of comparable similarity.

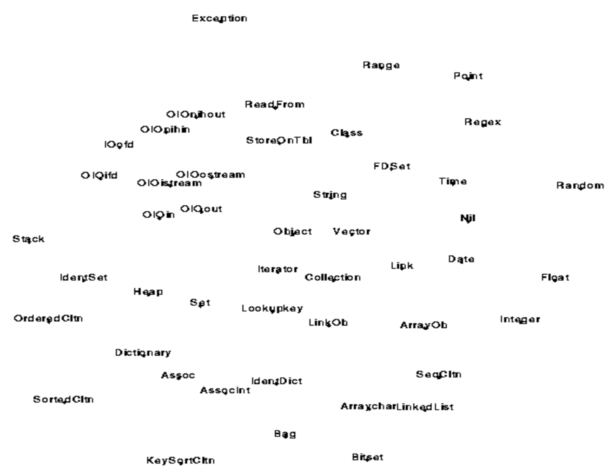

Figure 2: Sammon's Mapping of the NIHCL

On the other hand, the right part of Figure 1 gives the corresponding AC representation. Note that the overall structure of the text archive is clearly visible from the improved output visualization provided by the AC approach. The clusters are clearly separated from each other with the distances between both nodes as well as clusters providing information about their degree of mutual similarity. For example, the input/output classes again form a very strong and strictly separated, overlapping cluster in the upper right corner.[1] It is quite obvious that the previously described misleading perception resulting from the location of *Point, String* and *OIOofd* is clarified now. For a more detailed discussion we refer to [4].

The SM of the NIHCL is presented in Figure 2. The data points are arranged in a circular area with their relative locations resembling their distances in the high-dimensional space, representing a topology preserving mapping. To give an example, all classes dealing with input/output operations can be found grouped together in the upper left part of the mapping, the data types *Float* and *Integer* on the right of the mapping. However, as with the standard SOM representation, the detection of dissimilarities, i.e. cluster boundaries, is hardly possible without additional knowledge about the functionality of the classes. Although the distances between data points are not limited to a fixed grid distance, the differences in distance between related and non-related data points are too little to give clear evidence about the inclusion of data items in particular clusters. Thus, while providing information about the similarity of the input data, additional knowledge is required to detect the inherent structure of the text archive, i.e. to find the clusters and to understand their mutual

[1]The overlap of nodes is merely an inconvenience of the printed representation and can easily be resolved by zooming into the relevant area.

relationship.

In a nutshell, the enhanced visualization using AC allows the detection of clusters, cluster boundaries and mutual similarity, and thus the overall structure of the data set, in a very intuitive way while not interfering with the robustness of the standard SOM training process. Although similar to the SM as far as the type of representation is concerned, the principles to obtain the visualization are different. While SM tries to find a location in the 2-dimensional output space for every data point, the AC visualization originates in the self-organizing process of the SOM using its abilities to cope with noise and to generalize from the given input signals. As a benefit, the AC visualization provides a clear separation of clusters. Additionally, the information concerning the overall organization of the library, i.e. different topics, is readily accessible.

5. Conclusion

We have presented the application of an enhanced visualization technique for self-organizing maps called *Adaptive Coordinates* to aid in the detection and understanding of the overall structure of text archives. Text documents are transformed into a vector space representation on which a SOM is trained using an enhanced learning rule to mirror the movements of the map's nodes in a 2-dimensional output space. The resulting visualization depicts the inherent structure in the high-dimensional input space in a very intuitive way derived from the self-organizing process of SOM training. Mapping a query onto the SOM directs the user to the relevant part of the map which may be used as the starting point for convenient interactive exploration.

References

[1] Gorlen, K. E., NIH class library reference manual, National Institutes of Health, Bethesda, MD, USA, 1990.

[2] Kohonen, T., Self-Organizing Maps, Series in Information Sciences, Springer Verlag, 1995.

[3] Lagus, K., Honkela, T., Kaski, S., and Kohonen, T., Self-organizing maps of document collections, Proc. Int'l Conf. Knowledge Discovery and Data Mining (KDD'96), Portland, OR, 1996.

[4] Merkl, D., Exploration of Document Collections with Self-Organizing Maps, Proc. Symp. Principles of Data Mining and Knowledge Discovery (PKDD'97), Trondheim, Norway, 1997.

[5] Rauber, A., Cluster Visualization in Unsupervised Neural Networks, Diplomarbeit, Vienna University of Technology, Austria, 1996.

[6] Salton, G., Automatic Text Processing, Addison-Wesley, 1989.

[7] Sammon, J. W., A Non-Linear Mapping for Data Structure Analysis, IEEE Trans. on Computers, C-18(5), 1969.