A Self Organizing Map for Relation Extraction from Wikipedia using Structured Data Representations

Stephan Bloehdorn
Institute AIFB, University of Karlsruhe
D-76128 Karlsruhe, Germany
email: bloehdorn@aifb.uni-karlsruhe.de

Sebastian Blohm
Institute AIFB, University of Karlsruhe
D-76128 Karlsruhe, Germany
email: blohm@aifb.uni-karlsruhe.de

Abstract— **In this work, we will report on the use of self-organizing maps (SOMs) in a clustering and relation extraction task. Specifically, we use the approach of self-organizing maps for structured data (SOMSDs) (i) for clustering music related articles from the free online encyclopedia Wikipedia and (ii) for extracting relations between the created clusters. We hereby rely on the bag-of-words similarity between individual articles on the one hand but additionally exploit the link structure between the articles on the other.**

I. Introduction

The self-organizing map (SOM) [1] is a widely used neural-network approach which has been sucessfully employed in data visualization, dimension reduction and clustering tasks. Recently, extensions to the standard SOM model have been proposed that enable the analysis of complex structures like sequences, trees or graphs.

The general idea of this work is to use such an approach to automatically derive sets of documents – in our case articles from Wikipedia – with reference regularities of the type "articles on composers tend to reference articles of their major works" together with these references. This should be done in absence of a-priori knowledge whether articles feature musicians, songs or any other topic only with the help of untyped hyperlinks between the articles. Finding such regularities is an essential task in information extraction and ontology learning as it allows to identify classes, properties and relations relevant to the domain of interest – thus e.g. aiding the transition from the current Wikipedia to a "Semantic Wikipedia" as propsed in [2]. It is important to note that the definitions of classes, properties and relations interact: The class of composers can be reasonably defined as the class of people that stand in a particular relation to instances of the class of musical works.

We use the self-organizing map for structured data (SOMSD) [3], [4], an extension to the original SOM algorithm, to train the map to reflect not only similarity on the data level (i.e. textual similarity in our case) but also in the hyperlink structure. That is, proximity of two articles on the map allows not only to conclude that the articles have similar textual content, but also that they contain references to articles that in turn are close to each other on the map. The U-matrix clustering technique for SOMs [5] is applied to derive clusters based on map proximity. Rules like the one on composer articles referencing compositions can then be derived by observing the frequencies of references between cluster pairs.

II. Clustering with SOMSD and U-Matrix

In this section we shortly describe the main ideas behind SOMs and the U-matrix clustering technique and then focus on explaining the extension of the SOMSD approach.

Self Organizing Maps: The SOM is a powerful and intuitively understandable tool for unsupervised learning and data visualization [1] which combines vector quantization with a topological layout of the prototype vectors. SOMs allow for a mapping of high-dimensional input vectors onto a discrete output space (the "map") such that each region on the map represents an area of the input space. Preferably, this mapping should preserve the topology of the input space in the sense that local similarity of input patterns is reflected by proximity on the map. SOM training, i.e. the iterative adjustment of the prototype (weight) vectors to obtain a desired mapping, is done by successive presentation of all input patterns where each presentation includes the adjustment of weights to the presented pattern. Formally, the SOM can be defined as a set of nodes $\mathcal{N}$ (in neural network literature typically referred to as neurons) arranged on a grid which is typically a two-dimensional regular lattice, i.e. $\mathcal{N} \subset \mathbb{R}^2$. Each node is endowed with a weight vector $w(n) \in \mathcal{X}$ whereby $\mathcal{X}$ is the vector space of the input patterns, typically a real vector space i.e. $\mathcal{X} = \mathbb{R}^n$ with $n >> 2$. In this setting metrics can be defined for both, $\mathcal{N}$ and $\mathcal{X}$, which we will assume to be the standard Euclidean metric in $\mathbb{R}^2$ and $\mathbb{R}^n$ respectively. The mapping of input vectors $m : \mathcal{X} \to \mathcal{N}$ onto the map is defined as $m(x) = \texttt{argmin}_{n\in\mathcal{N}} ||w(n) - x||$, whereby $m(x)$ also referred to as *winner node*. During SOM training, the map is shaped through iterative presentation of input patterns each of which leads to slight adaptations of the map. We sketch the main ideas of the procedure:

1) Identify the winner node $m(x)$.
2) Update the weights of all nodes by a delta: $\Delta w(n) = h(n, m(x))\eta(x - w(n))$. Hereby, η is a learning rate parameter that may be adjusted during training and $h(n, m(x))$ is a function incorporating the distance of the neuron under consideration to the winning neuron, i.e. $\|n - m(x)\|$.

U-Matrix Clustering: The U-matrix (unified distance matrix) [5] is a SOM-based clustering technique that exploits the fact that when parts of the input space $\mathcal{X}$ are mapped onto the SOM, the area of the map representation correlates with the density of data samples from that part independent of the corresponding volume in $\mathcal{X}$. The U-matrix allows to detect such differences in density by assigning each node the sum of the distances of its weight vector to those of its direct lattice neighbours $u(n) = \sum_{n' \in neighbours(n)} ||w(n) - w(n')||$. Large U-values can be interpreted as borders between clusters, small U-values indicate homogenous neighbourhoods. Clustering is performed by grouping adjacent nodes whose U-values lie below a fixed threshold parameter (in analogy with a 3D interpretation of the U-matrix sometimes referred to as "water level").

Self Organizing Maps for Structured Data: The SOMSD is an extension to the SOM for scenarios where information is not only contained in the individual patterns but also in the order they are presented in [3], [4]. The order of the presentation can be used to encode sequences or paths in a directed graph on the input data. The SOMSD allows to take the context of a presented input pattern into account. This done by means of a recursive formula over the entries of a given sequence s during winner selection, a characteristic shared with other recursive SOM models. However, in contrast to other recursive SOM approaches as e.g. the Temporal Kohonen Map (TKM) proposed in [6], the SOMSD represents the sequence context only by the location of the winner neuron of the previous sequence item. As in the standard SOM model, each node $n \in \mathcal{N}$ is equipped with a *weight vector* $w(n) \in \mathcal{X}$; in addition, it is also equipped with a *context weight vector* in the vector space of the model, i.e. $w_c(n) \in \mathbb{R}^2$ for a two-dimensional Euclidean grid. The context weight vector denotes the preferred region of previous activation on the map. The distance of a presented pattern to a grid node thus becomes a mixture of two terms: the match of the neuron's weights and the current sequence entry on the one hand and the match of the neuron's context weight vector and the location of the previous winner on the map on the other. The new mapping of input vectors $m_c : \mathcal{X} \rightarrow \mathcal{N}$ onto the map is thus given by $m_c(x_i) = \texttt{argmin}_{n \in \mathcal{N}} \alpha ||w(n) - x_i|| + \beta ||w_c(n) - x_{i-1}||$, whereby x_{i-1} referrs to the pattern preceeding pattern x_i while α and β are coefficients to tune the influence of context. For the first pattern in a sequence, the predecessor is represented by `null` values and the context term is not used in winner selection like in the standard SOM. The actual adaptation of the SOMSD model works precisely as described for the standard SOM with the extension that the context weight vector is adapted in the same way as the standard weight vector. The SOMSD is intriguing due to two related properties: (i) the position of a pattern in a sequence becomes part of the characterisation of the pattern and (ii) the sequence structure can be easily reconstructed from the final map layout by simply using the context weight vector as a pointer to the node representing the preferred predecessor. Especially, the second aspect can be exploited to derive relations between *node clusters* built using the U-matrix approach by grouping the backward refrences within each cluster according to their respective cluster membership, thereby deriving an ordered list of preferred regions of predecessors for each cluster.

III. Experiments on Wikipedia

In this section way aim at presenting initial results of applying our approach to textual articles from the musical domain (and related areas) of Wikipedia, a free and collaboratively edited online encyclopedia. The experimental evaluation of our approach is targeted at discovering homogenous article clusters and relations between these clusters. While a large scale experimental application of SOMSD on our Wikipedia dataset is still ongoing we report on initial results and observations as well as problems experienced.

Wikipedia: The experimental evaluation is based on the Wikipedia version of March 26, 2006. Each article in wikipedia can be referenced uniquely by its title; as articles in Wikipedia typically describe individuals or concepts so that each article can be regarded as the representation of the entity in question. For simplicity, we will refer to the set of articles in question as $\mathcal{A}$ and to an individual article as $a \in \mathcal{A}$. Each Wikipedia article comprises (among others) the text of the article, a set of linked articles and a set of categories. The *text of the article* is a textual elaboration on the entity referred to by the article and will be denoted as $a.T$. By *links* of an article $a.L \subseteq \mathcal{A}, \forall a \in \mathcal{A}$ we refer to those Wikipedia articles that the article a links to by means of hyperlinks embedded in the text of the article. All Wikipedia articles further belong to at least one *Wikipdia category* which provides an alternative way of organization of the articles. We will denote the set of Wikipedia categories as $\mathcal{C}$ and the categories of a specific article as $a.C \subseteq \mathcal{C}$ with $|a.C| \geq 1, \forall a \in \mathcal{A}$.

Experimental Setup: We prepared a subset of 8,555 wikipedia articles. These articles were obtained by taking the set of 400 articles in the category *female american singers*[1] and all articles that can be reached from these via hyperlinks within two hops[2]. For texts and links we considered only the content of the first paragraph which is typically the most informative content. Term vectors were extracted using the standard preprocessing steps. [3] To cope with the inherent computational complexity of SOM training we reduced the dimensionality of the resulting feature vectors in two consecutive steps: from the overall collection of terms we retained only those occuring in at least 10 documents, resulting in a total number of 8,596 distinct term features, weighted using standard TFIDF; in a second step we performed dimension reduction by means of Latent Semantic Indexing (LSI) [7], retaining only the 200-dimensional approximations of the original feature vectors –

[1] `http://en.wikipedia.org/wiki/Category:American_female_singers`

[2] Note that the articles that can be reached from these seed articles need not be related to the musical domain (and typically aren't) – for example, about 8% of the articles crawled were related to dates (years and days), another large fraction was related to countries and other geographical articles.

[3] These being chunking, removal of the standard stopwords for English defined in the SMART stopword list and stemming using the Porter stermmer.

a common approach used with SOM training on textual data (cf. [8]). We generated a total of 250,643 two-item sequences required for the SOMSD training based on sample click sequences within the link structure. Note that this approach is comparatively simple and will result in a strong bias towards "hub" articles like dates and the like that are associated with a large number of incoming and/or outgoing links as these occur considerably more often within sequences than poorly linked articles.

Interpretation of Initial Results: As the computational complexity of SOM training is considerably higher than with other learning algorithms due to the repeated comparisons with all SOM grid neurons during winner selection a larger scale evaluation is ongoing. We here report on intitial results achieved on a 20x20 SOM grid taking into account 3 contexts, trained in 5 iterations. Note that one iteration involves all of the previously mentioned 250,643 two-item sequences resulting in an implicitely considerably higher training effect for the data weights. The α parameter was initially set to 0.6 and linearly decreased to 0.4 during the iterations, the correspondig parameter β was set to $1 - \alpha$ and thus evolved in the other direction. Note that as the α parameter scales the distance in a 200-dimensional space compared to a 2-dimensional space for β, the influence in winner selection of the data part is still considerably higher than the context part. The learning rate parameter η as well as the parameter σ which is used in the gaussian smoothing function $h(\cdot)$ were likewise linearly reduced from 0.2 to 0.02 and from 2 to 1 respectively.

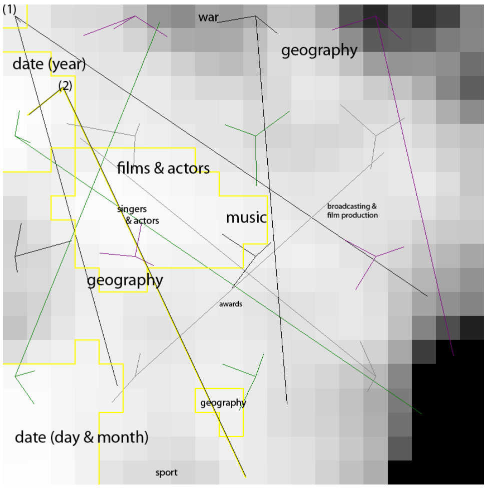

Fig. 1. Example 20x20 SOMSD using 3 contexts.

Figure 1 visualizes the resulting map. Lighter areas indicate smaller U-values corresponding to largely homogenous neighbourhoods. The resulting cluster structure indicates a number of regions of major topics. We have indicated the thematic correspondence within the map after manual inspection of winning neurons for the training sequences on the final map. The interpretation of the created context structures is significantly more subtle. We have plotted pointers to the coordinates referred within the context weights for a subset of 16 neurons with 5 neuron distance to each other in each dimension. The *date (year)* cluster in the upper left corner of the map is a good example of a structure that also exhibits the influence of the contexts. Consider the two neurons marked with *(1)* and *(2)*. While neuron *(1)* primarily attracts sequences where a year is linked to by another date page (day & month) whereas neuron *(2)* primarily favours a geographical context. Obviously, this elaboration is largely anecdotical and we also noticed that parts of the map are obviously not yet fully trained which is consistent with the observation that there were still major adaptations of the data and context weights taking place during the last iteration.

IV. Conclusions and Outlook

In this extended abstract, we have outlined the SOMSD approach as a means for generating topological maps for data items that are structured in a sequence-like manner. As the standard SOM, the SOMSD can be used for visualization purposes and for clustering, e.g. using the U-matrix approach. We have outlined the application setting of using SOMSD on Wikipedia articles whereby the sequence representation is determined by the hyperlink structure within Wikipedia.

Related Work: The usage of SOMs for visualization and clustering of textual documents — also in large numbers — has been pioneered in the context of the WEBSOM project [8]. The application of SOMs for sequence data in conjunction with U-Matrix clustering has been reported in [9]. Link mining has become a major research interest in the data mining community with a wide range of approaches beeing explored, see e.g. [10] for a recent survey.

Acknowledgment: This research was partially supported by the European Commission under the projects X-Media and SEKT. The expressed content is the view of the authors but not necessarily the view of the project consortia.

References

[1] T. Kohonen, Ed., *Self-organizing maps.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997.

[2] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer, "Semantic wikipedia," in *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, May 23-26, 2006*, MAY 2006.

[3] M. Hagenbuchner, A. C. Tsoi, and A. Sperduti, "A supervised self-organizing map for structured data," in *Advances in Self-Organizing Maps*, 2001, pp. 21–28.

[4] M. Hagenbuchner, A. Sperduti, and A. C. Tsoi, "A self-organizing map for adaptive processing of structured data," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 491– 505, May 2003.

[5] A. Ultsch, "Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series," in *Kohonen Maps*, E. Oja and S. Kaski, Eds., 1999, pp. 33 – 46.

[6] G. Chappell and J. G. Taylor, "The temporal kohonen map," *Neural Networks*, vol. 6, pp. 441–445, 1993.

[7] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[8] T. Kohonen, S. Kaski, J. Salojrvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE Transactions on Neural Networks*, vol. 11, pp. 574–585, 2000, second edition.

[9] M. Strickert, B. Hammer, and S. Blohm, "Unsupervised recursive sequence processing," *Neurocomputing*, vol. 63, pp. 69–98, 2005.

[10] L. Getoor and C. P. Diehl, "Link mining: A survey," *SIGKDD Explorations Newsetter*, vol. 7, no. 2, pp. 3–12, 2005.