A new hyperbolic visualization method for displaying the results of a neural gas model: application to webometrics

Shadi Al Shehabi[1] and Jean-Charles Lamirel[1]

1- LORIA - Campus Scientifique, BP 239
54506 Vandoeuvre-lès-Nancy Cedex - France

Abstract. The core model which is considered in this paper is the neural gas model. This paper proposes an original hyperbolic visualization approach which is suitable to be applied on the results of such a model. The main principle of this approach is to use a hierarchical algorithm in order to summarize the gas contents in the form on a hypertree in which information on data density issued from the original neurons (i.e. clusters) description space is preserved. An application of this approach on a dataset of websites issued from European universities is presented in order to prove its accuracy.

1 Introduction

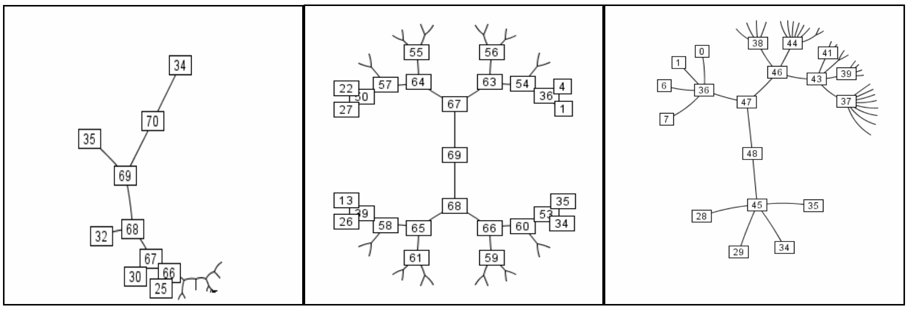
Data mining or knowledge discovery in database (KDD) refers to the non-trivial process of discovering interesting, implicit, and previously unknown knowledge from large databases. The efficiency of a neural gas (NG) model [1] for knowledge extraction has been highlighted in [2]. However, a large set of mining tasks, like webometrics, require an overall view of the analysis results where the interaction between the obtained classes is preserved. As soon as it should be applied to the results of a NG model, or even to the results of other data analysis methods producing multidimensional cluster descriptions, such kind of visualization represents a complicated problem that has not been yet solved. Hence, when linear projection methods, such as principal component analysis (PCA) or random mapping [3], are used for that purpose, the result is a significant loss of information. When non linear projection methods such as Sammon's Non-Linear Mapping [4], curvilinear component analysis (CCA) [5] or curvilinear distance analysis (CDA) [6] are used, there is no guaranty of a proper visualization of the neighborhood structure between clusters in the case of high dimensional data spaces. On its own side, hyperbolic visualization is known for its capability to cope with the problem of cognitive overload produced by the graph-based approaches [7]. Hence, it permits to visualize complex relationships between data through a focus and context mechanism. Up to now, hyperbolic visualization has mostly been used for solving data organization problems, like folder or hyperlink management. Although, hierarchically growing hyperbolic self-organizing map method has been proposed in [8] to cope with the limitation of the learned hierarchies of SOM visualization. Moreover, other approaches for clustering the results of the SOM have been proposed in [9]. However, all these latter methods do not cope with the limitation of the SOM model itself as compared to the NG model [1][10]. The original approach that is presented in this paper consists in combining NG based analysis with hyperbolic visualization. The hyperbolic visualization algorithm that will be used in our experiments is an algorithm that has been especially developed for this purpose. Its main advantage is to produce a hierarchical classification of the gas in which the information on data density issued from the original neurons (i.e. clusters) description space is preserved. Its secondary advantage is to indirectly highlight the original structure of the neighborhood between the neurons (i.e. clusters). The second section presents the hyperbolic visualization principles based on the MultiGAS model. A webometrics experiment is presented in the last section.

2 Hyperbolic hierarchical tree visualization

The hyperbolic tree visualization (hypertree), which has been invented by Lamping and Rao [11] is a *focus+context* visualization made for navigation in large tree-structured data. It uses hyperbolic geometry to provide its detailed and global views. The *focus+context* visualization is a property of the navigation which permits to provide a detailed view on a particular part of the data (*focus*) while still keeping a global view of the structure of the set (*context*). The properties of such a non-Euclidian geometry makes the hyperbolic space an ideal candidate for embedding large hierarchical structures of data [11]. The hypertree visualization uses the Poincaré disk for representing the hyperbolic space in two dimensions in the Euclidian space [12]. Hence it permits to obtain a *focus+context* representation of a tree. The branches of the tree turn into geodesics of the Poincaré disk, that is either diameters or circle arcs. The main advantage of such visualization is that the length contraction for nodes farther from the center of the disk allows users to keep a global view on the tree structure, even if the tree is really large. Moreover, the center of the disk offers a detailed view of the items in it [7].

The technique of the hypertree visualization has been created for representing arborescent hierarchical data. However, it can represent supplementary links between nodes. It permits a rapid and efficient navigation of a big data set. It also gives a good idea of a global organization of data. Hence, our original approach consists in combining gas analysis with hypertree visualization. The main principle of this approach is to use a hierarchical algorithm in order to summarize the gas contents in the form on an hypertree. The choice of the hierarchical algorithm is very important in this case, both for preserving information about the density issued from the original neurons (i.e. clusters) description space and for highlighting the original structure of the neighbourhood between the neurons (i.e. clusters) topography. Our proposed algorithm (see figure 3) behaves both as a density-based algorithm [13] and as a hierarchical algorithm, as soon as it preserves the density degree for each level of the hypertree: the lower the hierarchical level, the higher the density in the clusters. The original data (clusters) will be leaves of the hierarchical tree. The root of the hypertree represents itself the overall set of clusters. Moreover, a cluster which has many links (i.e. many sons), as compared to the other clusters of the same level, will represent a more dense area as compared to the other clusters. All leaves which have the root of the tree as direct parent cluster will be considered as outliers.

The figure 1 proposes a comparison between our hierarchical clustering algorithm and the more classical top-down hierarchical and reciprocal neighbors algorithms [14]. It shows that these two latter clustering algorithms do not preserve the information about the original data density and its topology in the description space.



(**a**) top-down hierarchical (**b**) reciprocal neighbors (**c**) density preserving

Fig. 1: Comparison between hierarchical clustering algorithms for hypertree building. The original dataset is a non homogeneous gas of 36 neurons issued from documentary data. The flags numbers represents the cluster indices (71 clusters for **a**, 70 clusters for **b**, 49 clusters for **c**).

3 Experimental results

The data used in our experiment are information about European University websites collected in work-package WP3 of the EICSTES project [15]. The investigated data set covers the Web sites of universities and research institutes in the 15 countries of the European Union before March 2004. The domain categorization of the websites is based on the UNESCO classification. The UNESCO code is a classification, which is used to allocate a scientific domain to of a Web site starting from its content. As the original dataset is too general, we have decided in a first step to focus our study on a specific thematic domain. The selection of data subset related to this domain is also based on UNESCO code. The 1203 UNESCO code that deals in a global way with computer science is used for websites extraction. A first context set of 2839 websites is selected in this way. In a second step, we have chosen to focus on German university websites as reference websites for our study. A second kernel set of 378 websites is selected this way. Hence, the study will more precisely focus on the relationships existing between German universities relatively to a European context.

The different information associated with the selected websites enables us to define 3 different viewpoints: (1) UNESCO codes (**Ucodes**), (2) German universities and their related cities (**Cities**) (3) Links coming from European universities to German universities (**In-Links**), Each viewpoint is represented in the form of a matrix including websites codes in its rows and specific viewpoint criteria in its columns. The matrices and the indexes will represent the input files for the basic clustering NG application. For each viewpoint an optimal gas is calculated. This gas is generated thanks to an optimization algorithm based on the quality criteria we have proposed in [16]. The optimization algorithm generated 50 neurons gases for each viewpoint.

Our experiment consists in building hyperbolic trees for visualizing the results of the gas analyses of two different viewpoints, that is to say the **Cities** viewpoint and the **UCodes** viewpoint. The goal of this is to highlight groups of interaction between cities (universities) and topics, respectively. The hypertree resulting of the **UCodes** viewpoint analysis is presented on figure 2.

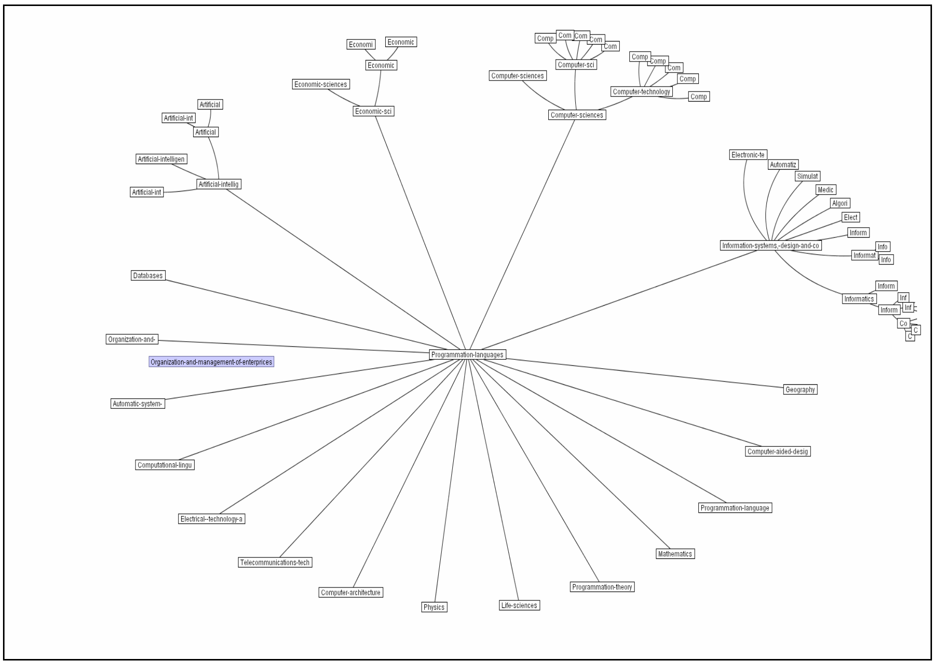


Fig. 2: Hyperbolic tree of the research topics managed in German universities. Topics that appear in neighbor sub-branches of the tree are strongly related one to another. As an example, **Simulation** and **Medical Science** are strongly related in German research.The **Information systems** topic (at the right of the tree) is also very rich because it includes many sub-branches. The central topic related to computer science in Germany is **Programmation languages**. Hence, it represents the root of the hyperbolic tree.

4 Conclusion

In this paper we have proposed a new approach that consists in combining gas analysis with hyperbolic visualization. Our first experiment of this approach has been carried out on a reference dataset of European websites that has been build up within the framework of the EISCTES project. Even if complementary experiments must be done, our first results are very promising. Indeed, they have shown that hyperbolic visualization represents a useful tool for interpretation of the results of a data analysis. Moreover, in the webometrics domain, this type of visualization can easily challenge the classical graph-based approach that often produces unmanageable results. This approach can also be applied to other data analysis methods that produce multidimensional cluster descriptions. In a near future, we also plan to apply it to the specific domain of link analysis. Moreover, the optimization of the number of levels in the hypertree generated by our method remains an open problem.

References

[1] T. Martinetz and K. Schulten, A "neural-gas" network learns topologies. In Kohonen, T., Mäkisara, K., Simula, O. & Kangas, J. (Ed.), Artificial neural networks (pp. 397-402). Amsterdam: North-Holland, 1991.

[2] J-C. Lamirel and S. Al Shehabi, Efficient Knowledge Extraction using Unsupervised Neural Network Models. In Proceedings of 5th Workshop On Self-Organizing Maps - WSOM 05, Paris 1 Panthéon-Sorbonne University, p.291-298, 2005.

[3] S. Kaski, Dimensionality reduction by random mapping. In Proc. Int. Joint Conf. on Neural Networks, volume 1, pages 413–418, 1998.

[4] J. W. Jr. Sammon, A Nonlinear Mapping for Data Structure Analysis. IEEE Transactions on Computers, vol. C-18, no. 5, pp 401-409, May 1969.

[5] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for linear mapping of data sets. 8(1):184-154, January 1997.

[6] J. A. Lee, A. Lendasse, N. Donckers and M. Verleysen. A robust nonlinear projection method. In European Symposium on Artificial Neural Network, pages 13-20, 2000.

[7] B. Bergé and C. Bouthier, Mathematics and algorithms for the hyperbolic tree visualization. Technical Report, A05-R-023, 15, September, 2003.

[8] J. Ontrup and H. Ritter, A hierarchically growing hyperbolic self-organizing map for rapid structuring of large data sets. In Proceedings of 5th Workshop On Self-Organizing Maps - WSOM 05, Paris 1 Panthéon-Sorbonne University, 2005.

[9] J. Vesanto and E. Alhoniemi. Clustering of the Self-Organizing Map. *IEEE Transaction on Neural Networks*, 11 (2):586-600, March 2000.

[10] S. Al Shehabi and J-C. Lamirel, Multi-Topographic Neural Network Communication and Generalization for Multi-Viewpoint Analysis, In Proceedings of International Joint Conference on Neural Networks (IJCNN'05). Montréal, 2005.

[11] J. Lamping and R. Rao, The Hyperbolic Browser : A Focus+Context Technique for Visualizing Large Hierarchies. Journal of Visual Languages and Computing, 7(1):33-55, 1996.

[12] M. Henle, Modern Geometries. Prentice Hall, 2001.

[13] M. Ester, H.P. Kriegel, H. Sander and X. Xu. A Density- Based Algorithm for Discovering Clusters in large Spatial Datasets with Noise. Proc. 2nd Int. Conf. KDDD. Portland, Oregon, p.1232-1239

[14] C. de Rham. La classification hiérarchique ascendante selon la méthode des voisins réciproques. Les Cahiers de l'Analyse des Données, 5(2): 135-144, 1980.

[15] EISTES project, European Indicator, Cyberspace and the Science-Technology-Economy System. IST-1999-20350, 1999.

[16] J-C. Lamirel, S. Al Shehabi, M. Hoffmann and C. Francois, New classification quality estimators for analysis of documentary information: application to web mapping. Scientometrics, 60 (3), 445-462, 2004.

Input: Dataset, n : maximum number of desired hierarchical level

Definitions: Each data of the initial dataset represents a leaf cluster
V_i : represents i^{th} level of the hierarchical tree; $0 \leq i \leq n$
M_i : represents the set of non agglomerative clusters, $|M_i|$ is the cardinal of this set
M_0: represents the initial dataset (leaf clusters); $|M_0|$ = number of original data.

/* Construct the new parent clusters in the different levels */

```
The matrix of distances between clusters of $M_0$ is constructed,
and two values are extracted of it: the maximal distance $Dmax$ and the minimal distance $Dmin$
Given a threshold $T$ such as: $T = ( Dmax – Dmin ) / n$
For each level $V_i$ Do
    $M_i = \emptyset$
    $Dmin = Dmin + T$
    If ($Dmin > Dmax$) Then Exit
    For each $C_j \in M_i$ Do
        Construct an initial parent cluster $C_j$*: $Parent(C_j)$.
        For each $C_k \in M_i \setminus \{ C_j \}$ Do
            If $|| C_j – C_k|| \leq Dmin$ Then
                Associate $C_k$ to $C_j$*
            End if
        End for
        The profile of $C_j$* is the average profiles of the associated son clusters.
        $M_i = M_i \cup \{ C_j^* \}$
    End for
    Procedure 1 ( $M_i$ )
    Procedure 2 ( $M_i$ )
    At this step $M_i$ contains all parent clusters and the clusters which have no parent
    cluster and the matrix of distances between the elements of $M_i$ is constructed.
End for
```

Procedure 1: elimination of repeated parent clusters

The repeated initial parent clusters in the new level (i.e. the initial parent clusters of the new level that share the same son clusters) are summarized into a single initial parent cluster.

Procedure 2: avoiding soft clustering

In this case the son clusters which belong to more than one initial parent cluster will be associated to one single parent cluster:

1- For each initial parent cluster, recalculate its profile without the shared associated son clusters.
2- Associate each son cluster to the initial parent cluster whose profile is the nearest: said initial parent cluster will become the final parent cluster.
3- If the initial parent clusters of a shared son cluster have the same distance to the said son cluster then associate the shared son cluster to any of them in arbitrary way.
4- If an initial parent cluster has only one son cluster then eliminate this parent cluster.

Fig. 3: Hierarchical density-preserving clustering algorithm.