Artificial Neural Network Technology: for the Classification and Cartography of Scientific and Technical Information

Xavier Polanco, Claire François, Jean-Pierre Keim

Artificial Neural Network Technology: for Classification and Cartography of Scientific and Technical Information.

Xavier POLANCO1, Claire FRANCOIS1 and Jean-Philippe KEIM1

1Institut de l'Information Scientifique et Technique (INIST)
Centre National de la Recherche Scientifique (CNRS)
2 allée du Parc de Brabois
54514 Vandoeuvre les Nancy cedex - France

This paper describes the implementation of multivariate data analysis: NEURODOC applies the axial k-means method for automatic, non-hierarchical cluster analysis and a Principal Component Analysis (PCA) for representing the clusters on a map. We next introduce Artificial Neural Networks (ANNs) to extend NEURODOC into a neural platform for the cluster analysis and cartography of bibliographic data. The ANNs tested are: the Adaptive Resonance Theory (ART 1), a Multilayer Perceptron (MLP), and an associative network with unsupervised learning (KOHONEN). This platform is intended for quantitative analysis of information.

1. INTRODUCTION

Informetrics' general fonction is the analysis of information (Polanco 1996). The aim of our work is to perform this general fonction by computer using automatic data cluster analysis algorithms and the representation of the clusters so generated in the form of maps. We apply this approach to the domain of Scientific and Technical Information (STI).

The STI analysis involves the use of three indicator types or levels: keywords as content indicators, of the knowledge conveyed by the documents; clusters as indicators of the topics or the centres of interest contained in the information (articles, authors, institutions, journals); map as strategic indicator of the relative position of the topics in the knowledge space covered by the documents analysed.

These techniques (cluster analysis and representation) belong to the mathematical domain of statistical analysis known as multivariate data analysis. The statistical analysis of the information (informetrics) can be completed with the aid of neural networks (models which are essentially non-linear and threshold-driven).

In the design and development of informetric techniques, one path to be explored is the application of connectionist models which use learning mechanisms to classify and represent data. Our interest in artificial neural algorithms lies in the links which exist between data analysis and the connectionist approach. ANNs and statistical methods of multivariate data analysis have much in common. Connectionist methods can be considered as non-linear data analysis methods. PCA, k-means or dynamic cloud clustering methods correspond to unsupervised neural methods; regression and linear discriminant analysis are special cases of supervised neural methods (Lebart et al. 1995; Cheng & Titterington 1994; Ripley 1994; Lelu 1991, 1993).

Our first studies of the application of ANNs to informetrics have resulted in the definition of the axial k-means method of cluster analysis. This method is inspired from the neural formalism of Kohenen's model is an application of a modified version of Oja's winner-takes-all type of learning rule (Lelu 1993). This method is implemented in NEURODOC (Lelu & François 1992a, 1992b).

In this article, we first present the approach used by NEURODOC for the analysis of the information (section 2). We then present the first stage of a project in which ANNs are used to develop NEURODOC into an ANN platform (Section 3).

2. NEURODOC

The three main components of NEURODOC are: [1] Cluster analysis which groups the documents by cluster, and therefore also the authors, their affiliations and the journals in which they were published. This cluster analysis is achieved using the axial k-means method. [2] A factor representation of topics (or clusters) identified above based on the PCA. [3] A hypertext interface

generator for PC (under WinHelp) and for Macintosh (under Hypercard) which provides the user with a user-friendly interface with the map, the topics and the documents themselves.

2.1 Classification

As its name suggests, the axial k-means method is a variant of the k-means method developed by MacQueen (1967). MacQueen's algorithm belongs to the family of "moving centre" cluster analysis algorithms. Alain Lelu (1993), the author of the "axial k-means" method, presents it as a synthesis of factor analysis and cluster analysis, drawing on the neural formalism of Kohonen's model, that is to say, the associative projection model. This method considers the collection of bibliographic references as a cloud of points in a multidimensional space where each dimension corresponds to a keyword. Characteristically, it represents clusters as vectors pointing towards areas of highest density. The usual non-hierarchical clustering techniques represent the *k* clusters found according to their centre of gravity; the axial k-means define the *k* clusters found by *k* half-axes passing through the origin of the multidimensional space, or *k* unit vectors pointing in the direction of the half-axes.

The position of the *k* half-axes is initialised either at random or by the first *k* documents. We then calculate the square projections $y_i(k)$ of each normalized document *i*, upon the *k* half-axes thus defined, by calculating the scalar products of the normalized document *i* with the unit vectors of the *k* half-axes. Each document is attached to the cluster *k* where the projection *y(k)* on the axis *0Ak* is maximal. To take this attachment into account in the adaptive form of the algorithm, the position of the axis is immediately recalculated. In the iterative form it is recalculated after all the documents have been treated. By successive iterations the axes are repositioned and stabilise in the high density areas of the data cloud, thus forming a strict classification of the documents.

To obtain overlapping clusters, a "typicality threshold" is defined: while a document belongs to the cluster to which it was associated during the final pass, it can also belong to a different cluster if the value of its projection upon this second axis is greater than the threshold. A document can thus belong to several clusters if the values of its projection upon the corresponding axes are greater than the threshold.

For example, the value of the projection of the document *i* upon the various clusters is greatest for the axis $Ak\ (y_i(k))$ while the value of its projection upon the axis $Ak'(y_i(k'))$ is less than the threshold. The document *i* thus belongs uniquely to the cluster *k*. The value of the projection of the document *ii* upon the various clusters is greatest for the axis *Ak* and the value of its projection upon the axis *Ak'* is greater than the threshold. The document *ii* belongs thus to both clusters *k* and *k'*.

Documents belonging to a given cluster can be ranked according to the value of their projection upon the axis which represents the cluster. For example, the projection of the document *ii* upon the axis *Ak(yii(k))* is greater than that of document *i (yi(k)).* This order corresponds to a decreasing order of "typicity" of the documents compared with the "ideal" type of the cluster which is a fictitious document positioned exactly on the axis in the multidimensional space.

By using the values of the components of the cluster unit vector, we can define a partition of keywords from the documentary corpus in the same way. As for the documents, the partition thus created can lead to overlapping clusters. A keyword can belong to several clusters and keywords are ordered according to the decreasing pertinence to the ideal cluster type. The weighting used to determine the value of pertinence is used to bring out specific (or typical) keywords for the cluster, that is to say, frequent in the cluster and rare in the documentation overall.

This algorithm, which can be parameterized with the maximum number of clusters desired and the threshold of co-ordinates of the documents and keywords on the axes, is used to construct clusters of a particular type. [1] These clusters overlap because a document or a keyword can belong to several clusters at once. [2] The documents or keywords are ranked according to their degree of resemblance to the ideal cluster type.

2.2 Cartography

A cluster of documents corresponds to a topic, a homogeneous sub-set of the information contained in the documentary corpus being studied. The PCA of the group of clusters in the multidimensional space is used to determine a plane causing the least deformation to the clusters cloud. All the points representing clusters are then projected upon this plane to create the global topic map. On such a map, two topics which appear at opposite ends of the axis represent clusters defined by keywords which are as different as possible (see figure 2).

2.3 Automatic Hypertext Generation

The STI which has thus been structured by classification and cartography is then organized in hypertext. The hypertext is used to obtain an overview of the documentary corpus from the global map and then to gain access to pertinent information organised by topic (clusters). Selecting a topic allows one to see its description in the form of four lists [1] keywords, [2] titles, [3] authors, [4] journal titles.

In 1995, INIST used NEURODOC for producing five thematic data files based on PASCAL and FRANCIS databases related to "Pain and pain treatment", "Physiology of central nervous system receptors", "Cosmetics", "Natural energy sources", and "Human resources". These data files have been distributed in French on diskettes. These applications also exist on a CD-ROM version (dmcontact@inist.fr). In addition, these applications can be accessed and searched with HENOCH system over the Internet (Grivel et al. 1997).

3. MOVING TOWARDS AN INTEGRATED PLATFORM

In order to develop NEURODOC into a platform allowing the application of ANN to bibliographic data, we have initiated an applied research project in collaboration with the CORTEX group of the RFIA Team, common to CRIN-CNRS and INRIA-Lorraine (Nancy, France; http://www.loria.fr). This section describes the results of the first stage of this project: the application of unsupervised formal neural networks to the domains of classification and cartography:

- To clusterize bibliographic data, we needed a network with unsupervised neural clustering analysis capabilities similar to axial k-means clustering. Our raw data are represented as binary vectors with 0 and 1 values. The ART 1 meets these requirements. The MLP produces supervised clustering being more similar to the Discriminant Factor Analysis.
- To map topics, (i.e. to place topics produced by the clustering phase into a map), we used two ANNs: a self-assoaciation MPL which, in this case, is similar to the PCA approach and KOHONEN self-organizing map (SOM) (Kohonen 1982, 1995).

The aim of this project is to show how the implementation of the various neural methods can assist with the analysis of the information provided. This is done with a view to its reusability and insertion in NEURODOC to produce a neural information analysis platform.

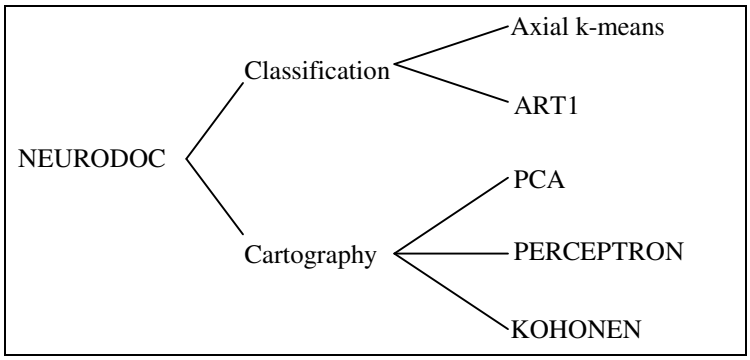

Figure 1: Evolution of NEURODOC towards an ANN platform

3.1 Content of utilized data

We tested these algorithms on a corpus of bibliographic data from INIST's PASCAL database. These data contain all references pertaining to cosmetics from this database as published in 1995. This corpus includes 276 records, 95% of which are derived from 85 journals. NEURODOC clustering produced 20 topics as shown on the map in the figure 2. This corpus is available for evaluation.

3.2 Cluster Analysis by Adaptive Resonance Theory (ART)

The ART 1 used here comes from the work of Carpenter and Grossberg (1988). The cluster analysis is performed by using the keywords as indicators of the contents of the bibliographic data. The data are represented in the form of a matrix where the rows correspond to the documents and the columns to the keywords. This matrix contains 0s and 1s which represent the presence (1) or absence (0) of a keyword in a document. Such a matrix is principally constructed of 0s. The aim

of ART 1 is to associate a set of codes to a group of data. These codes determine the data clusters as according to their resemblance. The main interest of ART 1 is its capacity to create as many codes as necessary. These codes will ultimately be used to define the clusters. The most important thing to bear in mind in trying to understand ART 1 is that only the 1s contain any information. The resemblance is calculated using the following formula:

$$\frac{P_i \cdot E^k}{\left\| E^k \right\|^2} \geq \rho$$

Where P_i and E^k are two vectors of the same dimension representing a code and a data respectively; ρ is the vigilance parameter, a real number between 0 and 1. $P_i * E^k$ gives the number of bits set to 1 in both P_i and E^k; $\|E^k\|^2$ gives the number of bits set to 1 in E^k. It appears then that the bits set to 0 are not considered at all in calculating the resemblance between two vectors. This implies that if a data and its code have 90% of their 0s in common, this does not mean that they will necessarily be associated.

This resemblance depends on two factors: ρ, the vigilance parameter which determines the minimum resemblance between a data and its code in terms of the number of bits set to 1, and β, a coefficient which gives added weight to codes having the most number of 1s when associating data to a code.

The first stage in the study of ART 1 was to verify the influence of these two factors on the number of clusters produced. The effect of β turns out to be very small. For our corpus of 276 documents, the minimum number of clusters obtained is about 76. In order to reduce the number of clusters, we have removed the codes where only one bit is set to 1 and redistributed the documents in the remaining clusters. This results in the appearance of documents which cannot be classified. These documents have no keywords in common with the remaining codes. So as not to lose them, we have created a separate cluster containing all these documents. We have then developed a method of merging clusters so as to end up with 18 clusters.

Assessment: Although the number of clusters produced is higher compared with the number of documents to be classified, the main problem is to evaluate the pertinence of these clusters as indicators of the research topics. The codes corresponding to the clusters contain few keywords. In order to evaluate them and compare them with the results obtained with the axial k-means, it is necessary to characterise the clusters better by defining an order of pertinence for the documents associated with the clusters, as well as for all the keywords which index the documents. ART 1 is simple and has limited capacity. It can only treat binary entries. However, ART 2 network can treat vectors with continuous values allowing a cluster analysis of normalized documents. We plan to study both these ART networks.

3.3 Cartography using MULTILAYER PERCEPTRON and KOHONEN NETWORK

As well as being a means of visualization, maps also represent an analytical method insofar as they can be used to evaluate the relative position of topics in a multidimensional representation space.

The use of the PCA in the domain of bibliographic data is confronted with the fact that the data are very multidimensional. The representativeness of the first two axes when measured as an inertia percentage is always weak (about 20 to 30%). There are always a number of isolated topics in the multidimensional space which are misleadingly positioned on the map by the orthogonal projection.

In order to find a better system to represent the clusters generated by the axial k-means method, we have tested a MULTILAYER PERCEPTRON (MLP), and a KOHONEN network. A number of improvements have been made to MLP algorithm following the advice given by Jodouin (1994). Kohonen's algorithm was defined according to Jodouin (1994) and Hertz et al. (1991). The input data provided to these networks are represented by a matrix where the rows correspond to the clusters (fewer than 50 in general) and the columns to keywords (several hundred in general).

3.3.1 Three layers PERCEPTRON

We have not used the PERCEPTRON for its ability to make associations but for its spatial projection capabilities. We want the output signal to be the same as the input while passing through an intermediate layer with only two neurones. This network is thus required to reduce a n-

dimensional vector into a two-dimensional one and then to reconstruct a n-dimensional vector. This topology is used to obtain two co-ordinates: the activation values of the central neurones for each topic. The learning level of the network is known due to the Mean Quadratic Error (MQE) which is calculated periodically. This MQE corresponds to the Euclidean distance between the input and output vectors. Its evolution has three phases: a relatively long plateau, a step-wise descent, and a final stabilization. The MQE is given by the following formula:

$$MQE = \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}(S_{i,j} - T_{i,j})^2 \quad (1)$$

where N is the number of clusters to place; M is the dimensions number of vectors representing clusters; $S_{i,j}$ is the value taken by the j^{th} neurone for the i^{th} example; $T_{i,j}$ is the value expected for the j^{th} neurone in the i^{th} example.

Assessment: We have seen whether the grouping and separation of the topics reflect the contents of the corpus by submitting the maps with their hypertexts to expert in documentation. We have defined a three-stage evaluation protocol. [1] The Euclidean distances between the vectors of all the clusters are calculated. [2] The position of the topics is verified in the following manner: for each cluster studied, the positions of the others is considered correct; in this way we verify that the cluster is well centred in comparison with those nearest to it (criterion 1) and a long way from those which are most different (criterion 2). A topic is well positioned if those two criterions are respected, and moderately well placed if only one criterion is respected. [3] By a global consideration of the results of the second step, we have been able to evaluate the quality of each map.

In this way we have seen that those maps obtained by the PERCEPTRON (figure 3) were clearly superior to those obtained by the PCA (figure 2). The distances on these maps are not proportional to the distances calculated in N-space but the oppositions are virtually all respected.

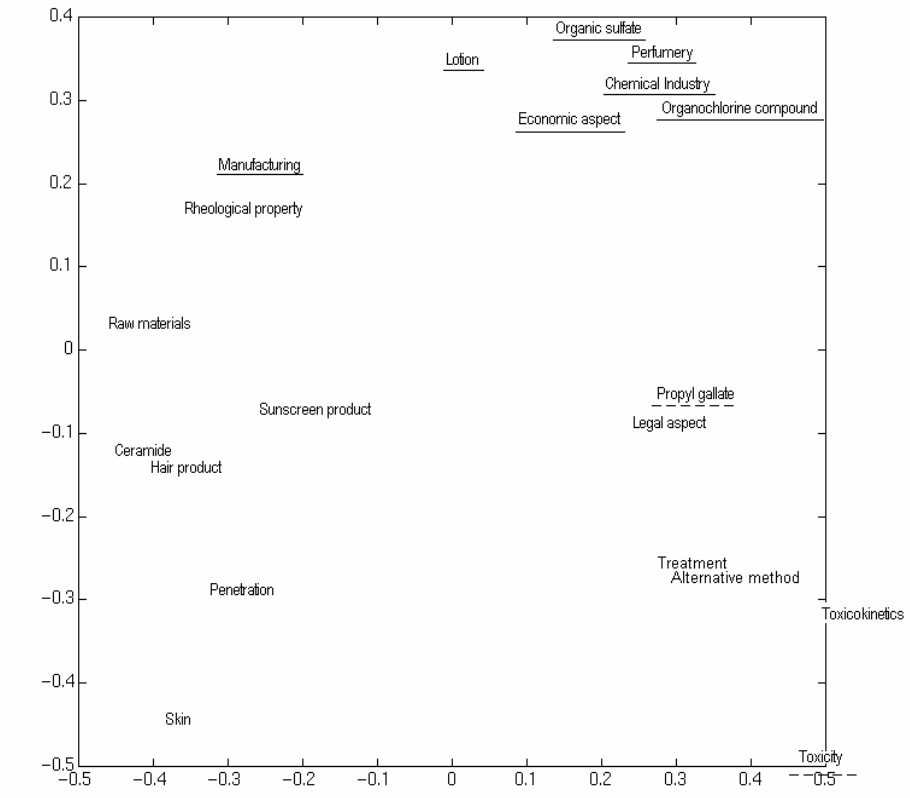

Figure 2: Sample of a topic map obtained by PCA on a corpus in cosmetics. The topics underlined by a solid line are badly positioned while those underlined by a dotted line are moderately well placed.

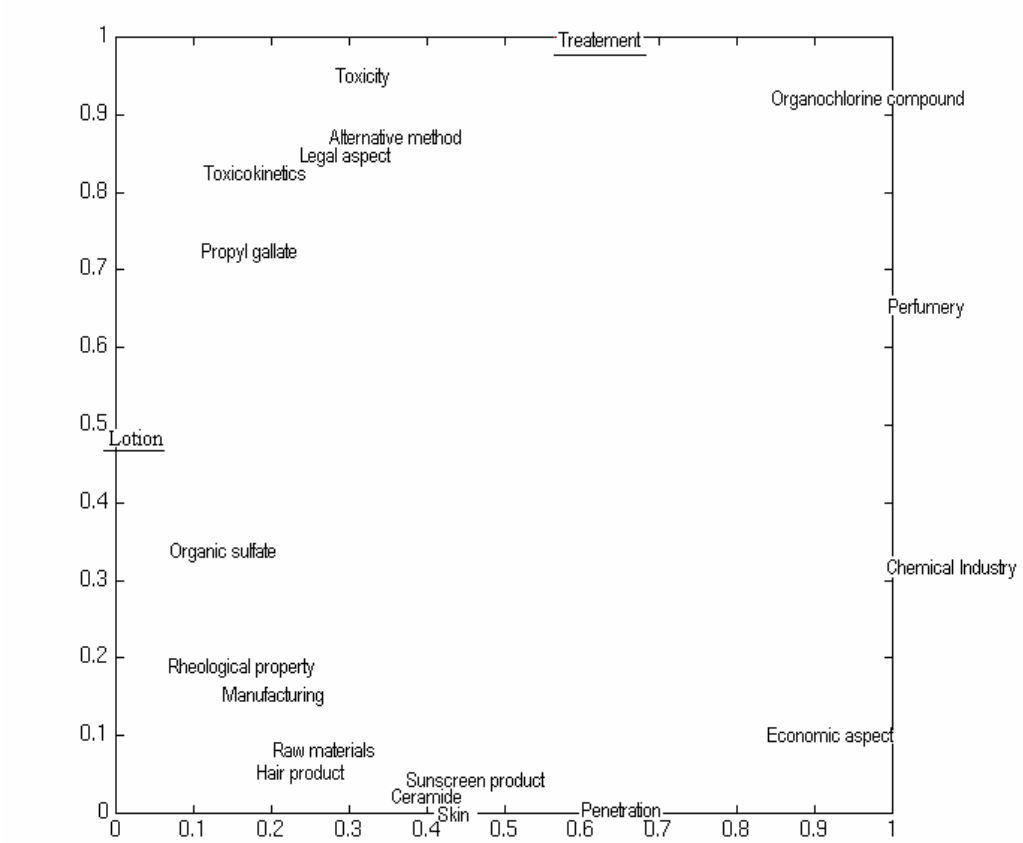

Figure 3: Example of a topic map obtained by the three-layer PERCEPTRON.

3.3.2 KOHONEN Network

The KOHONEN network has a two-layer topology consisting of an input layer and a output grid and learns in an unsupervised manner. This structure allows us to place the n-dimensional data directly in the plane represented by the grid. The learning follow up is achieved in the same way as for the PERCEPTRON, i.e. following the evolution of the MQE.

For the PERCEPTRON, we used the vectors representing the clusters to the input but, in the case of KOHONEN, to improve its performance, we presented the vectors representing the keywords, in other words the transposed matrix. We obtained a keyword map where the co-ordinates of topics on the map are calculated in the following manner: the position of the topic is the barycentre of the co-ordinates of all the keywords making up the topic.

Assessment: We note that the KOHONEN method produces results more quickly than the PERCEPTRON**.** However, the final MQE is higher than the three-layer PERCEPTRON. The map evaluation protocol is the same as that used for the three-layer PERCEPTRON.

As we have already mentioned, we are planning to test the KOHONEN model in its unsupervised clustering on competition-based learning aspects and its mapping capabilities. In this respect, Campanario (1995) used a KOHONEN self-organizing map algorithm in its work on journal-to-journal citation data. He noted that this algorithm has the following advantages compared to the multidimensional methods: [1] an improved representation of the asymmetric form of the cross-citation matrix among journals, and consequently [2] a better understanding of the nature of interrelations between journals.

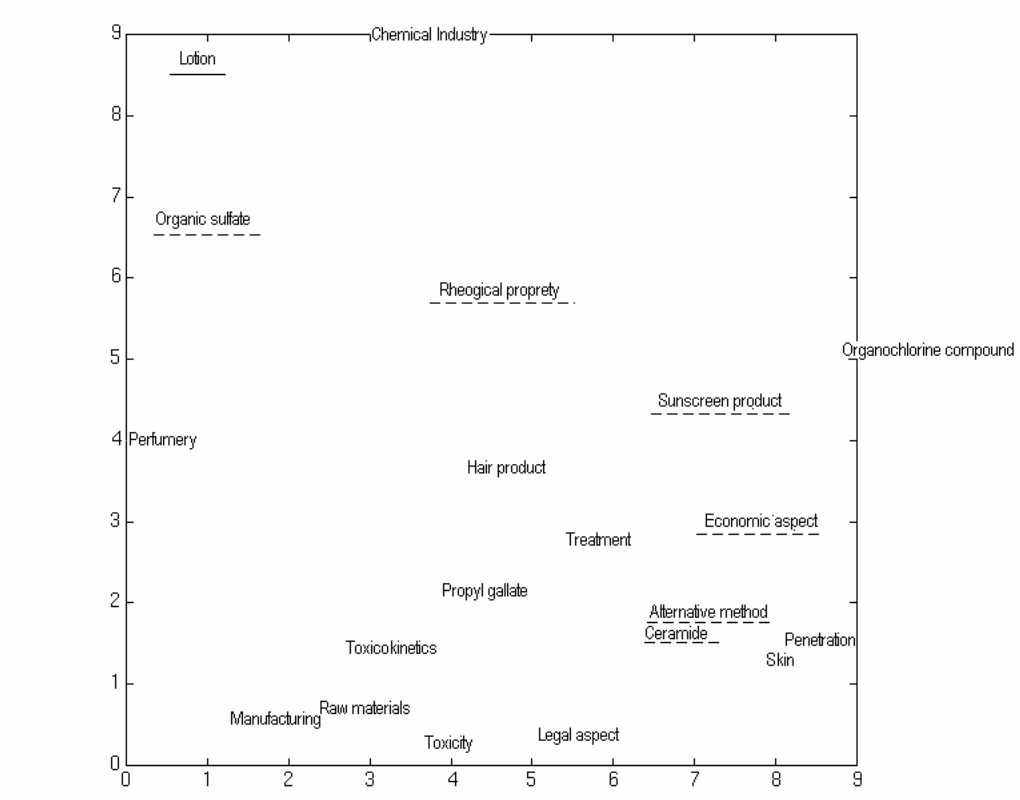

Figure 4: Example of a topic map obtained using the KOHONEN method.

3.3.3 Evaluation and Perspectives of the two neural methods

It is necessary to find a statistical evaluation methodology to measure the effect of the ANNs in the cartographic representation of topics. This method could be based on the Euclidean distances between topics and give values to their closeness and/or separation, measure the occupation of the map, and how well the order of proximity is respected. Once such a tool has been created, we will be able to study the modifications brought about by the transposition of data. Then, we could envisage optimising the map using a simulated annealing method. In terms of the implementation of the PERCEPTRON and KOHONEN programmes, it would be interesting to alter the stop conditions. Up until now, we have had two possibilities, either to wait for stabilisation in the MQE or to impose a certain number of iterations. In the light of the results, it may be interesting to use the following formula:

$$MQE = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(S_{i,j} - T_{i,j})^2 \quad (2)$$

where N is the number of clusters to place; $S_{i,j}$ is the value taken by the j^{th} neurone for the ith example; $T_{i,j}$ is the value expected for the j^{th} neurone in the i^{th} example. If the MQE is less than a fixed threshold, stop. Otherwise, continue. We think that it would be easy to calculate a stop threshold as a function of the number of examples. The aim of this calculation would be to impose a maximum tolerated average distance between the input and the output.

4. CONCLUSION

In this paper, we have attempted to present the preliminary results and assessments of a research project oriented towards the application of neural algorithms for the cluster analysis and cartography of the STI. We now have an ANN platform at our disposal. The first stage of the implementation was realised between April and July 1996 (Keim 1996). The results that we have shown (section 3) must be followed by more statistically-rigorous evaluation tests in order to compare [1] the results obtained by the different cluster analysis methods, axial k-means, ART 1

and ART 2, classification by self-associating KOHONEN maps, and [2] the maps obtained by the statistical (PCA and MDS) and neural methods (PERCEPTRON and KOHONEN).

We believe that we have filled a gap in informetrics concerning the use of ANNs already exploited for information retrieval (Doszkocs et al. 1990). By applying to quantitative studies of science and technology, methods used by science itself in many domains (see, for example, the bibliometric study of Van Raan and Tijssen in 1993 on Neural Network Research), we are creating a back-propagation movement known here as the "science of science", and which is the source of scientometrics (Price 1963).

Finally, let us to point out that our aim is not limited to use ANNs for quantitative studies of bibliographical data in STI, but above all to develop a computerized informetric platform based on ANNs. Such a platform should be able to combine, or to apply the different statistical and neural approaches consecutively, according to the user's analysis and mapping strategies.

We believe that this design is one of the most promising ways to obtain implicit, unknown, but potentially useful information from a database. This is called today "data mining" or "knowledge discovery in databases". Our ongoing research is oriented towards a such direction. Note in passing that the technologies designed in order to analyze, evaluated and apply information to define strategies, can be called intelligence technologies. The NEURODOC platform can be seen as an example of this sort of technology.

ACKNOWLEDGMENTS

We would like to pay tribute to our deceased colleague Pascal Blanchet for his very valuable contribution to this project. We are grateful to Marie Christine Lunel, Isabelle Pignone, and Isabelle Clauss for their helpful scientific content expertises of the map building by ANNs techniques. This work was supported by a grant 93.K.6461 from DISTB-MENESRIP.

REFERENCES

Campanario, J. M. (1995) Using neural networks to study networks of scientific journals. *Scientometrics* 33 (1): 23-40.

Carpenter, G.A.; Grossberg, S. (1988) The ART of Adaptative Pattern Recognition by a Self-Organizing Neural Network. *Computer* 21: 77-88.

Cheng, B.; Titterington, D.M. (1994) Neural Networks: A Review from Statistical Perspective. *Statistical Science* 9 (1): 2-54.

Doszkocs, T.E.; Reggia, J.; Lin, X. (1990) Connectionist Models and Information Retrieval. *Annual Review of Information Science and Technology* 25: 209-260.

Grivel, L.; Polanco, X.; Kaplan, A. (1997) A Computer System for Big Scientometrics at the Age of the World Wide Web. *Proceedings of the Sixth International Conference on Scientometrics and Informatrics*. June, 16-19 1997, Jerusalem, Israel, 131-142.

Hertz, J.; Krogh, A.; Palmer R.G. (1991) *Introduction to the Theory of Neural Computation*. MA, Addison-Wesley Reading.

Jodouin, J.F.(1994) *Les Réseaux Neuromimétiques*. Paris, Hermès.

Keim, J-Ph. (1996) *Mise en oeuvre de réseaux neuronaux dans un logiciel d'analyse de l'information scientifique et technique*. Technical report (INIST-ISIAL), Nancy, France.

Kohonen, T. (1982) Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern*. 43: 59-69.

Kohonen, T. (1989) *Self-Organisation and Associative Memory.* Berlin, Springer Verlag.

Kohonen, T. (1995) *Self-Organizing Maps*. , Berlin , Springer.

Lebart, L.; Morineau, A.; Piron, M. (1995) *Statistique Exploratoire Multidimensionnelle.* Paris, DUNOD.

Lelu, A. (1991) From Data Analysis to Neural Networks: New Prospects for Efficient Browsing through Databases. *J. of Information Science* 17: 1-12.

Lelu, A. (1993) *Modèles Neuronaux pour l'Analyse de Données Documentaires et Textuelles.* Doctoral thesis from the Université de Paris 6.

Lelu, A.; François, C. (1992a) Information Retrieval based on a Neural Unsupervised Extraction of Thematic Fuzzy Clusters. *Conférence "Les Réseaux Neuro-Mimetiques et leurs Applications"*, 2-6 November, Nîmes, France.

Lelu, A.; François, C. (1992b) Hypertext Paradigm in the field of Information Retrieval: a Neural Approach. *Proceedings of the 4th ACM Conference on Hypertext*, 30 November - 4 December, Milan, Italy.

MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proc. 5th Berkeley Symp. Math. Proba.*: 281-297.

Polanco, X. (1996) La notion d'analyse de l'information dans le domaine de l'information scientifique et technique. *Conférence INRA - Information scientifique et technique*, 21-23 October, Tours, France.

Price, D. de S. (1963) *Little Science, Big Science*. New York, Columbia University Press.

Ripley, B.D. (1994) Neural Networks and Related Methods of Classification, *J.R. Statist. Soc. B*. 56 (3): 409-456.

Van Raan, A.F.J.; Tijssen, R.J.W. (1993) The Neural Net of Neural Network Research. An Exercise in Bibliometric Mapping. *Scientometrics* 26 (1): 109-192.