

Visual Exploration of Astronomical Documents

Soizick Lesteven, Philippe Poinçot, Fionn Murtagh¹

*CDS, Observatoire astronomique de Strasbourg, 11, rue de l'Université,
67000 Strasbourg, France*

Abstract. The CDS bibliographical map is a tool for organizing astronomical text documents into a meaningful map for exploration and search. The system is based on the Self Organizing Map (SOM) algorithm that automatically organizes documents into a two-dimensional grid so that related documents appear close to each other and general topics appear in well defined area.

After the determination of optimal parameters for the SOM's learning process, we have developed a graphical WWW interface which allows the visualization of the document distribution. It shows the localization of documents related to given topics (keyword queries). The map is clickable and provide links to the documents. Recent developments include detailed map of small areas, full text indexing, automatic labeling, ...

Some applications will be presented. One map is available for interactive use on the Web (<http://simbad.u-strasbg.fr/A+A/map.pl>).

1. Introduction

The continually increasing quantity of textual data requires constant effort in order to update storage and access methods so that the totality of information is easily accessible. Scientific publications are no exception. Astronomy is a good example in view of the enormous mass of data collected by modern satellites and large ground-based facilities, and the numerous scientific articles which result from such data.

The Centre de Données astronomique de Strasbourg (CDS) collects and organises different types of astronomical information (Genova et al. 2000). In particular, the CDS offers on-line access to some bibliographic data.

This article presents ongoing work on the CDS bibliographical maps. These maps are a tool for automatically organizing collections of astronomical text documents and for displaying them in order to facilitate the mining and retrieval of information (Poinçot 1998, 1999). The map is clickable and provides links to the on-line documents. The system is based on the Self-Organizing Map, or Kohonen Map (Kohonen 1995). In the article we present new developments, including detailed map of small areas, full text indexing, and automatic labeling.

¹School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland

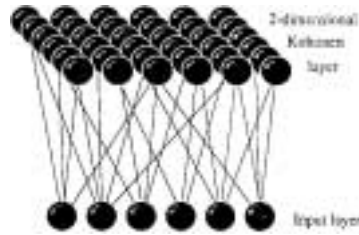


Figure 1. The topology of the Kohonen self-organizing map network.

2. The Self-Organizing Map (SOM)

The SOM is an algorithm used to visualize and interpret large high-dimensional data sets. These maps are one of the artificial intelligence techniques, and more precisely an unsupervised neural network. The topology of the Kohonen SOM network is shown in Figure 1. This network contains two layers of nodes: an input layer, and a mapping (output) layer in the shape of a two-dimensional grid. The input layer acts as a distribution layer. The number of nodes is equal to the number of attributes associated with the input. Each node of the mapping layer also has the same number of attributes. Thus, the input layer and each node of the mapping layer can be represented as a vector which contains the number of features of the input. The mapping nodes are initialized with random numbers. Each actual input is compared with each node on the mapping grid. The “winning” mapping node is defined as that with the smallest Euclidean distance between the mapping node vector and the input vector. The value of the mapping node vector is then adjusted to reduce the Euclidean distance. In addition, all of the neighboring nodes of the winning one are adjusted proportionally. After all of the input is processed (usually after tens of repeated presentations), the result should be a spatial organization of the input data organized into clusters of similar regions.

3. Application to Bibliographic Classifications: Creation of Bibliographical Map

We applied the method to the classification of articles published in *Astronomy and Astrophysics* in the period 1994 up to now (9450 articles). The attributes were based on the bibliographic keywords. We kept only the 269 keywords that appear in at least five different articles. The documents characterized in this way constitute a set of 9450 stimuli to be applied to the network. Learning through 50 iterations (heuristically determined) gives good results for the principal map (15×15 units).

We have adapted the use of the SOM to our own needs. Documents located at a map edge have neighbors at the other side of the map. It is then possible to reconfigure the map without losing the similarity of closely clustered documents. At the end of the training of a map, the number of documents assigned to each node is known. Because it is much easier to visualize the colours of an image than a matrix of numbers we transformed the table of numbers into an image. For this image the colour scale indicates qualitatively the number of documents per

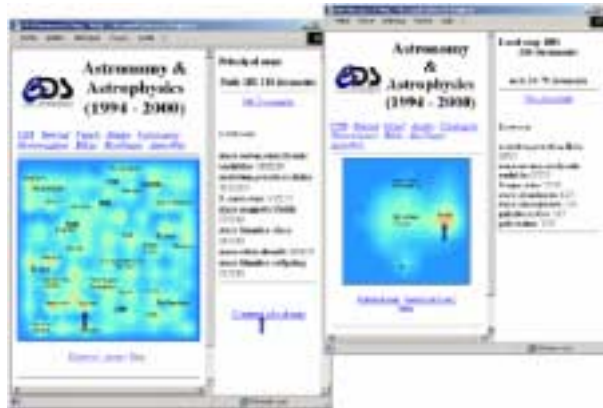


Figure 2. The secondary maps.

node. The map is manually labeled to locate on it the different themes associated with the nodes. The user can select one node of the map, by clicking on the picture, to obtain some information about the articles located in it: the number of documents and the keywords describing them appear on the right side of the interface. The user can also access the article content (title, authors, abstract) and all the facilities provided by the CDS bibliographical service (including a link to ADS and to the on-line full paper when available). The map can also be reached by keyword queries and bibcode queries (standard definition of a document). One map is available for interactive use on the World-Wide Web (<http://simbad.u-strasbg.fr/A+A/map.pl>).

4. Recent Developments

4.1. Secondary Maps

The size of the SOM map has a strong influence on the quality of the classification. The smaller the number of nodes in the network, the lower the resolution for the classification. The fraction of documents assigned to each node correspondingly increases. It can then become difficult for the user to examine the contents of each node when the node is linked to an overly long list of documents. However, there is a practical limit to the number of nodes: a large number means long training. A compromise has to be found. That is why, for each “over-populated” node of this map, termed the *principal map*, another network, termed *secondary network* or *map*, is created and linked to the principal map. Each secondary network is trained using the documents associated with the corresponding node and its surrounding nodes of the principal map (Figure 2). In this way, a map is created with as many nodes as necessary, while keeping the computational requirement under control.

4.2. Full Text Indexing

To overcome the limitations due to the representation of the documents by keywords, we use all textual information present in the documents (title, abstracts,

keywords, ...). We built our own full text indexing tool. All the words are kept and counted, empty words (without meaning) and the less frequent ones are eliminated. A truncature (Porter 1980) is applied to reduce morphologic variants of a word to a single index term. The program allows to keep word associations (2 or 3 words). The first applications show that the two-word associations are meaningful. It is now possible to create bibliographical maps with a full text indexing and to take into account documents for which keywords are not available. It seems that the full text indexing is on one hand noisier due to the automatic indexing (the context is not taken into consideration) but on the other hand more precise because it does not rely on a limited list of keywords.

4.3. Automatic Labeling

For map interpretability, the different themes associated with document/node assignments have to be indicated. Although our maps have a relatively limited number of units, it is important to automate the annotations. Because it is impossible to characterize all nodes without overlapping annotations it is preferable to select a limited number of nodes for characterization. The selection is made according the density peaks and the cross-positions of two peaks. When the position is defined, the process examines the words associated with the documents assigned to the peak and writes the most frequent one beside the peak. When different words have the same occurrence, the automatic choice (the first one) is not automatically the best one. Furthermore, our system makes it possible to classify the words in various categories. It is then possible to annotate the map according to the different classes, corresponding to a precise application.

5. Conclusions

We developed a visual and efficient system to explore astronomical documents based on the SOM. Other data collections, notably catalogue information, have already been processed in the same way, and a cartographic user interface tool has been set up to allow catalog selection (<http://vizier.u-strasbg.fr/>).

Now that new tools have been developed, a number of new applications are possible. Creation of a bibliographical maps for the SIMBAD objects, for an ADS query, for a colloquium organisation, ...

References

- Genova, F., et al. 2000, *A&AS*, 143, 1
- Kohonen, T. 1995, *Self Organizing Maps* (Berlin: Springer-Verlag)
- Poinçot, P., et al. 1998, *A&AS*, 130, 183
- Poinçot, P., 1999, PhD Thesis "Classification et recherche d'information bibliographique par l'utilisation des cartes auto-organisatrices, applications en astronomie"
- Porter M. F. 1980, *Program*, 14(3), 130