A spatial user interface to the astronomical literature

Soizick Lesteven, F. Murtagh, P. Poinçot

ASTRONOMY & ASTROPHYSICS
SUPPLEMENT SERIES

Astron. Astrophys. Suppl. Ser. **130**, 183–191 (1998)

A spatial user interface to the astronomical literature

P. Poinçot[1], **S. Lesteven**[1], **and F. Murtagh**[2,1]

[1] Observatoire Astronomique, 11 rue de l'Université, F-67000 Strasbourg, France
[2] Faculty of Informatics, University of Ulster, Londonderry BT48 7JL, Northern Ireland

Received September 26; accepted December 1, 1997

Abstract. We recall the properties of the Kohonen self-organizing feature map (SOM or SOFM), and explain how such maps can be used for information retrieval. We present an application to a bibliographic database. Our neural net can contain more than one level when necessary, which allows users to modify its spatial configuration. It is available for interactive use on the World-Wide Web (http://simbad.u-strasbg.fr/A+A/map.pl). The interface that we have designed for browsing in the documentary database will be explained in detail.

Key words: astronomical databases: miscellaneous — publications, bibliography — techniques: miscellaneous

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 Strasbourg Data Centre (CDS) has the role of collecting and organising different types of astronomical information (Egret et al. 1995; Genova et al. 1996). In particular, the CDS has the charge to build and offer online access to some bibliographic data from *Astronomy and Astrophysics* and its Supplement Series.

This article presents work ongoing at CDS on the use of an artificial intelligence technique applied to the classification of scientific articles. This technique is Kohonen's self-organizing feature map (SOM; Kohonen 1995). The way this particular type of neural network works will be described first. Then we will see how we have built a system appropriate for the classification, retrieval and consultation of documents. A description of the use of this search system, which we could call a *bibliographic map*, will conclude this article.

Information retrieval by means of "semantic road maps" was first detailed by Doyle (1961). The spatial metaphor is a very powerful one in human information processing. As we will see in this paper, the spatial metaphor also lends itself well to modern distributed computing environments such as the World-Wide Web (WWW). Semantic road maps are not necessarily based on neural networks. In fact there are quite varied approaches to visual information retrieval interfaces (Pörner 1995). Nor is the SOM approach the only neural network approach which could be used: for example Zavrel (1996) favours an adaptive grid (growing cell structures network). However we find the Kohonen approach to be a highly effective method. This method is validated on real data in this paper. This method is computationally tractable and produces stable results of high quality. It is compatible with widely-available software – for instance, no Web browser add-ons are required contrary to the case of other visual interfaces.

2. Kohonen maps

A Kohonen map is usually taken as a two-dimensional scene in which the observations are classified such that those which share related characteristics are located in the same zone of the map.

As an example of the application of the SOM method, Honkela et al. (1995) created a map of the fairy-tales of the Brothers Grimm. Word triplets were examined in this work, following removal of very frequent or very rare words. A triplet was used to provide a context for the middle word. The words in the text were classified without a priori syntactic or semantic categorization. The output map shows three clear zones where the set of nouns and verbs are separated by a more heterogenous zone constituted by prepositions, pronouns and adjectives.

A Kohonen map shares many properties with statistical (i) clustering methods, and (ii) dimensionality-reduction methods, such as principal components

Send offprint requests to: S. Lesteven

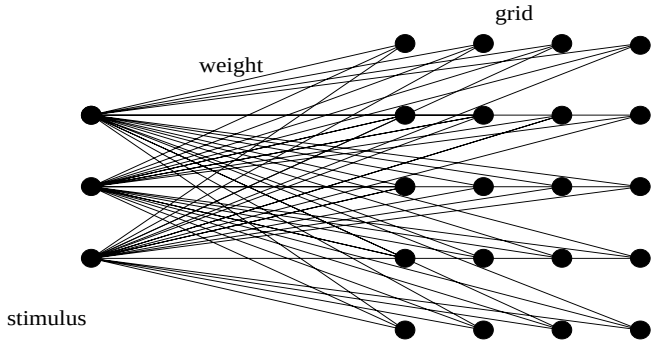

Fig. 1. The Kohonen self-organizing feature map network scheme

analysis and correspondence analysis. Murtagh & Hernández-Pajares (1995) present a range of comparisons illustrating these close links.

2.1. Creating a SOM Map

SOM maps are a particular type of neural network or pattern recognition method known as unsupervised learning. An SOM map is formed from a grid of neurons, also usually called nodes or units, to which the stimuli are presented. A stimulus is a vector of dimension d which describes the object (observation, entity, individual) to be classified. This vector could also be a description of the physical characteristics of the objects/stimuli. In this work, it will be based on characteristics such as the presence or absence of keywords attached to a document. Each unit of the grid is linked to the input vector (stimulus) by means of d synapses of weight w (Fig. 1). Thus each unit is associated with a vector of dimension d which contains the weights, w.

2.2. The algorithm

The grid is initialized randomly.

2.2.1. The learning cycle

The learning cycle consists of the following steps.

1. Present an input vector associated with a *stimulus* to the grid.
2. Determine the *winner* node. This is the unit for which the associated vector is the most similar to the input vector.

$$\|\text{input} - \text{node}_{\text{winner}}\| = \min_i \|\text{input} - \text{node}_i\|$$

3. Modify the weights w_i of the winner node, as well as those nearby, such that the associated vectors (the weight vectors) are as similar as possible to the input vector (p_i) presented to the grid.

$$w_i(t+1) = w(t) + h(r,t)(p_i - w_i(t)) \quad \text{if} \quad i \in \text{neighbourhood}$$

$$w_i(t+1) = w_i(t) \quad \text{if} \quad i \notin \text{neighbourhood}$$

where

$$h(r,t) = \alpha(t)v(r)$$

$\alpha(t)$ is the learning coefficient and $v(r)$ is the neighbourhood function.

4. Decrease the size of the neighbourhood of the winning nodes (the zone which contains neurons allowed to undergo modification).
5. Decrease the learning coefficient, $\alpha(t)$, which controls the importance of the modifications applied to the weight vectors.
6. Halt the learning when the learning coefficient is zero. Otherwise present another stimulus to the grid.

2.2.2. The neighbourhood function $v(r)$

The modification of vectors associated with the units is carried out in different ways depending on the position of the nodes with respect to the winner unit. The winning node will be the one whose vector will be potentially subjected to the most modification, while the more distant units will be less affected. The function $v(r)$ will be maximal for $r = 0$ and will decrease when r increases (i.e. at increasing distance from the winner node). In this work we used a linear function ($v(r)$) which allows for inhibition of nodes which are distant from the winner node.

2.3. The results

SOM maps allow the classification of objects for which we do not have a priori information. Once the map is organised (i.e. once the learning has been accomplished), each object is classified at the location of its "winner". The use of a grid containing fewer units than objects to be classified allows the creation of density maps. For such a grid, the distances between the objects to be classified can be related to axis interval lengths.

2.4. Treatment of boundaries

The nodes at the edges of an SOM map are to be treated with care, since they do not have the usual number of neighbouring units. To insure the stability of the map configuration during successive learning cycles, we took the view that a node at a map edge has neighbours at other extremities of the map. Thus our map is a flattened version of a sphere, allowing for wrap-around, as represented in Fig. 2.

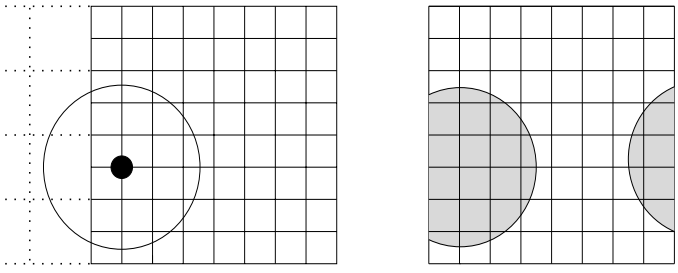

Fig. 2. Left: the neighbourhood of the winner extends over the boundary of the grid. Right: the gray zone shows the neurons whose weights are to be modified during training

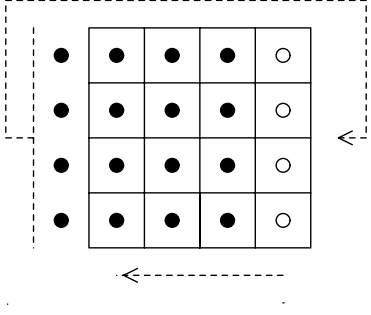

Fig. 3. Detail view of reconfiguring the neuron grid

As long as the neighbourhood radius is not large, wrap-around neighbourhoods do not cause any problems. If, however, the neighbourhood were to extend back into itself, this would imply ambiguities where certain units would be considered as close neighbours in one direction, while being very distant in the opposite direction.

The user can obviously modify the map at will by moving the rows or columns horizontally or vertically. A row pushed out beyond an edge will reappear at the opposite side (Fig. 3). This reordering can be of benefit when an interesting zone is located near the map extremity.

2.5. Influence of number of network nodes

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.

One possible strategy which we use to face this trade-off problem is to create another layer of maps (Luttrell 1989; Bhandarkar et al. 1977). The first network "of reasonable size" (to be further clarified below) is built and is trained using all stimuli. This network may be too small for an acceptable classification and some nodes may be linked to too many documents. Then, 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 of the principal map. The training of secondary maps is thus based on a limited number of documents, and not on all of them.

In this way, a map is created with as many nodes as necessary, while keeping the computational requirement under control.

2.6. Training time

Most of the computational requirement is due to the determination of the winner node corresponding to each input vector and to the modification of the vectors associated with neighbouring units of the winner nodes. The number of operations to be carried out is directly linked to the size of the network and to the number of stimuli:

$$\text{Determining the winners: } \sim N_{\text{stim}}.N_{\text{unit}}$$

$$\text{Updating of vectors: } \sim N_{\text{stim}}.\sum_{R=1}^{R_{\max}} R^2$$

where N_{stim} et N_{unit} are respectively the number of stimuli and the number of units of the network. $R_{\max}$ and R are the values of the radius of the neighbourhood at the start (fixed) and in the course of the learning (varying). The second formula is somewhat simplified because we have assumed a linear decrease in R from $R_{\max}$ to 1, while the learning distribution might be different.

Let us compare now the "classical" and "two-layer" approaches. Let us take, as an example, a primary map of dimensions 15×15, for which *each node* is linked to a secondary map of dimensions 5×5. The size of the corresponding "classical" map is thus 75×75.

Determining the winning nodes:

- In the case of the 75×75 map, the time required for determining the winner nodes is proportional to $5625N_{\text{stim}}$.
- For the two-layer system, the time is reduced to

$$\underbrace{225N_{\text{stim}}}_{\text{primary map}} + \underbrace{25\sum_{k=1}^{225} N_k}_{\text{secondary maps}} = (225+25)N_{\text{stim}}$$

  where N_k is the number of documents classified within the node k of the principal map.

The two-layer method is therefore about $5625/250 = 25$ times faster than the classical method, in this case, at this step.

Updating the units:

– In the case of the classical 75×75 map, the time required for updating the values of the winning nodes is proportional to

$$N_{\text{stim}} . \sum_{R=37}^{1} R^2 = 17575 N_{\text{stim}}$$

– In the case of the two-layer method, we have

$$\underbrace{N_{\text{stim}} . \sum_{R=7}^{1} R^2}_{\text{primary map}} + \underbrace{\sum_{k=1}^{225} (N_k . \sum_{R=2}^{1} R^2)}_{\text{secondary map}}$$

which gives

$$N_{\text{stim}}(140 + 5).$$

The results presented here are contrary to the claim made in Zavrel (1996) that the Kohonen map does not scale well. We have also found convergence properties, and in particular stability, not to give rise to undue difficulties. Other experiments concerning the stability of the results, described in Murtagh & Gopalan (1997), are in agreement with this finding.

To conclude, the association of a Kohonen secondary map with each node of a principal map, or with the over-populated nodes only, allows a high-quality classification of the stimuli with considerable improvement in the training time.

2.7. Weighting of index terms

The use of secondary maps for bibliography classification indicated early on the need for changing our mode of defining the document descriptors. We recall that the set of stimuli to be presented to the secondary map is the result of an earlier training. Therefore these stimuli are all more or less the same, and only differ by a small number of the descriptors present in the set. Often the documents are clustered around an over-burdened node. This is clearly not desirable since relevant information can be added by the secondary map in this case.

This behaviour is explained by the training principle: the descriptors or the index terms associated with the majority of the stimuli are those which occur most frequently, since the vectors associated with network nodes are modified to resemble the input vectors. During training, the corresponding components of these vectors take the largest values and therefore have the strongest weight. Finally, the classification of the documents is dominated by their relation to these majority descriptors. It is therefore not surprising that many documents are found associated with a few nodes.

The solution to this problem is to bypass the binary stimuli-vectors and to allow for vectors with non-integer components. Weights are used to specify the importance to be accorded to different descriptors. We used for this a simplification of the weighting method described by Salton (1989). This method uses n_i, the number of occurrences of descriptor number i in the set of documents to be classified in the secondary map; and N, the number of documents in this set to be classified in this way. The weight of descriptor i is then given by

$$w_i = \ln\left(\frac{N}{n_i}\right).$$

The input vectors are then normalized to prevent penalizing the stimuli associated with rare descriptors. The most frequent descriptors are downgraded, having small weights, which tend towards zero if they are present in all documents. By weighting the documents in this way, we were able to avoid the effect of accumulated stimuli at the same node. This assumes of course that the various stimuli are genuinely different.

3. Application to bibliographic classification: Creation of a bibliographic map

We applied the method just described to the classification of articles published in *Astronomy and Astrophysics* in the period 1994 to 1996 (3325 articles). The descriptors were based on the bibliographic keywords. For this journal, and some others in astronomy, there is a uniquely defined list of keywords. A necessary preliminary phase was to homogenize them (since they were assigned by one of the Editors but not in a completely systematic fashion). We kept only the keywords that appear in at least 5 different articles, so we limited our descriptors to the 250 most frequent. The documents characterized in this way constitute a set of 3325 stimuli to be applied to the network. Learning through 20 iterations (heuristically determined) gives good results for the primary map (15×15 units) requiring about 1 hour of processing on a Sparcstation 10 (dependent on other users). For the secondary maps, the learning time is much shorter, since fewer documents are processed (200 at most), and the network has a smaller dimensionality (25 units).

When the training of all the maps, principal and secondary, is finished, the next task is to make these maps accessible to the user.

3.1. Density maps

At the end of the training of a map, the number of documents assigned to each node is known. We therefore have a table of numbers. Because it is much easier to visualize the colours of an image than a matrix of numbers we transformed it into an image. For this image the colour scale indicates qualitatively the number of documents per node. The primary map is of dimension 15×15, and each of the secondary maps is 5×5. These images are then

scaled up by a factor of 40 (determined by aesthetics and most common Web browser default window sizes). This transformation uses a linear interpolation since otherwise the map would have clear discontinuities.

3.2. Indexing the maps

For map interpretability, the different themes associated with the document/node assignments have to be indicated. Although our maps have a relatively limited number of units, while in comparison the maps proposed by the team of T. Kohonen (WEBSOM 1997) have about 8 times more neurons, it is still impossible to characterize all nodes without overlapping annotations. Therefore it is preferable to select a limited number of nodes for characterization.

These nodes are selected from the frequent occurence of a keyword, which is written on the map. This was done manually, but could later be automated. The strategy is as follows:

- determine the density peaks;
- examine the keyword associated with the documents assigned to the peaks;
- write the most significant one beside the peak.

4. User interface to the maps

The HTML language allows an interface to be created which allows access to maps and their assigned documents, and in particular allows use of interactive or "clickable" images. At the start of his/her query, the user finds a principal bibliographic map, which is a density map corresponding to a principal Kohonen map. Aided by information inscribed on the image presented, he/she can then select a node of the network. Then, the keywords most frequently found in the associated documents are listed on the side of the map, as well as the corresponding number of documents. If the keywords are of interest to the user, he/she can click to access the appropriate secondary map when the node of the primary map is over-populated. If there is no over-population, the list of documents is immediately accessible.

The secondary maps are of the same form as the primary one and are used in the same way.

Note that when the list of articles is presented, the user can access the bibliographic services of the Strasbourg Data Centre (Fig. 7 below; Egret et al. 1995; Genova et al. 1996) in order to get the complete reference and abstract, astronomical object names cited in the article with links to the SIMBAD database for each object, and to tables from the paper when available in the CDS catalogue service, and in some cases to the complete text of the article.

The user interface is either via the cartographic display, or directly via keywords. The latter may be found of greater help to the user. Thus two interfaces are simultaneously supported: the clickable map, or the clickable selection of the keywords of interest.

Table 1. Index term lists for Novae units

Node Novae: main map
stars:novae,cataclysmic variables
accretion,accretion disks
X-rays:stars
stars:binaries:eclipsing
stars:magnetic fields
stars:binaries:close
stars:white dwarfs

Node Novae: secondary map
stars:novae,cataclysmic variables
accretion,accretion disks
X-rays:stars
stars:atmospheres
stars:binaries:general

5. An example

To illustrate and analyze the organisation of the Kohonen map applied to the articles from *A&A* between 1994 and 1996 (Fig. 4), we will study in detail a small area of that map concerning novae.

Looking at the map around the node labelled Novae, we find the general keywords: *Nova, neutron stars, close binary stars, ISM.* By definition, a nova is a cataclysmic variable, therefore a star that suddenly increases in brightness. Observations have demonstrated that novae are close binary stars of which one component is a white dwarf. When the companion star evolves and expands to fill it, material streams towards the white dwarfs, forming an accretion disc around it.

The neutron stars are formed in supernova explosions and are observed as pulsars. A stellar remnant of three solar masses or more will collapse into a black hole rather than a neutron star.

A quick comparison between the nova area with its specifications, and the general definition of a nova, are in agreement. The list of keywords attached to the Nova node, listed in Table 1, describes correctly a nova.

Let us look now at the articles attached to the Novae unit. We retrieve 77 documents from *A&A* (Fig. 5). This set of documents is too large for reading all the abstracts. This is why a more detailed map is proposed. The secondary map (Fig. 6) shows in fact a distinction between the documents concerning *white dwarfs, novae* and *binary stars.* At the new Nova node on the secondary map, the keywords are more limited but more precise. They are

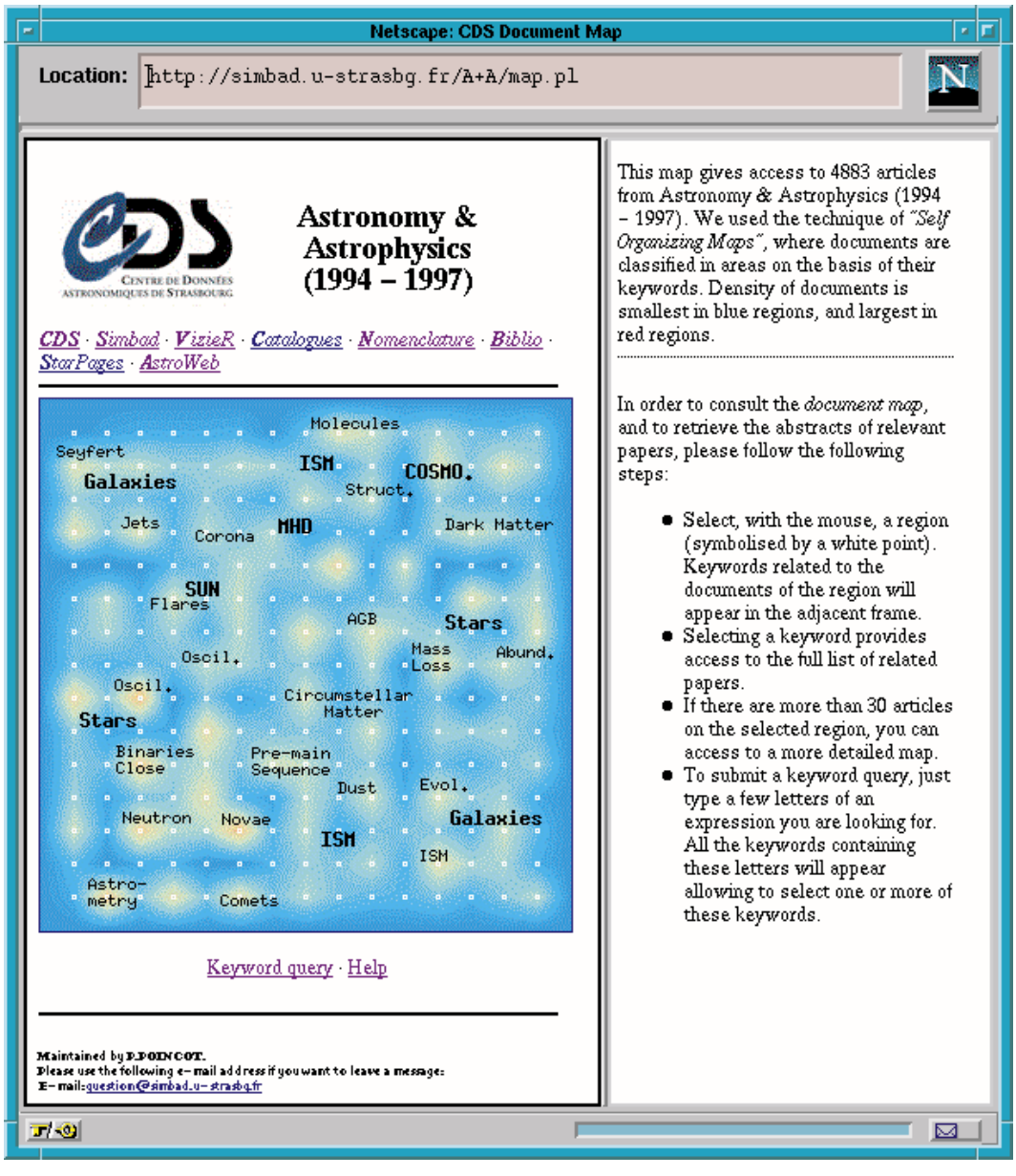

Fig. 4. View of the principal map

listed in Table 1. At this unit, we retrieve 38 documents; the first 7 documents are listed in Fig. 5.

All the documents deal with cataclysmic variables, nova-like stars, X-ray stars. Therefore they are clearly right within the subject.

Now, have a look at the nodes surrounding the Nova one on the main map. Above, we find documents concerning *accretion disks, close binary stars* and *white dwarfs*, to the left, *pulsars* and *neutron stars*, to the right, *supernovae* and *molecular data* and below, *comets* and *meteors*. There are logical links between these subjects.

Starting from the close binary stars node, we move into a wider stellar area described by different peaks defined by these main keywords: *stars:fundamental parameters, stars:oscillations* and *stars:variables, stars:activity, X-rays:stars*, and so on. Furthermore, we can move into the solar area by means of a link between the stellar oscillations and the solar oscillations.

In conclusion, the relationships presented in the map have been validated in most cases. We have also studied the evolution of annual maps, over the period of three years considered in this work. Little change was noted and we do not show these annual maps. Of course some well-known events relating to fashion or appearance of some astronomical phenomenon may well be reflected in our bibliographic data.

An alternative access to bibliography maps is implemented, allowing the user to select one or several keywords and to vizualize their positions on the maps. It is then possible to retrieve and select the relevant papers at these locations.

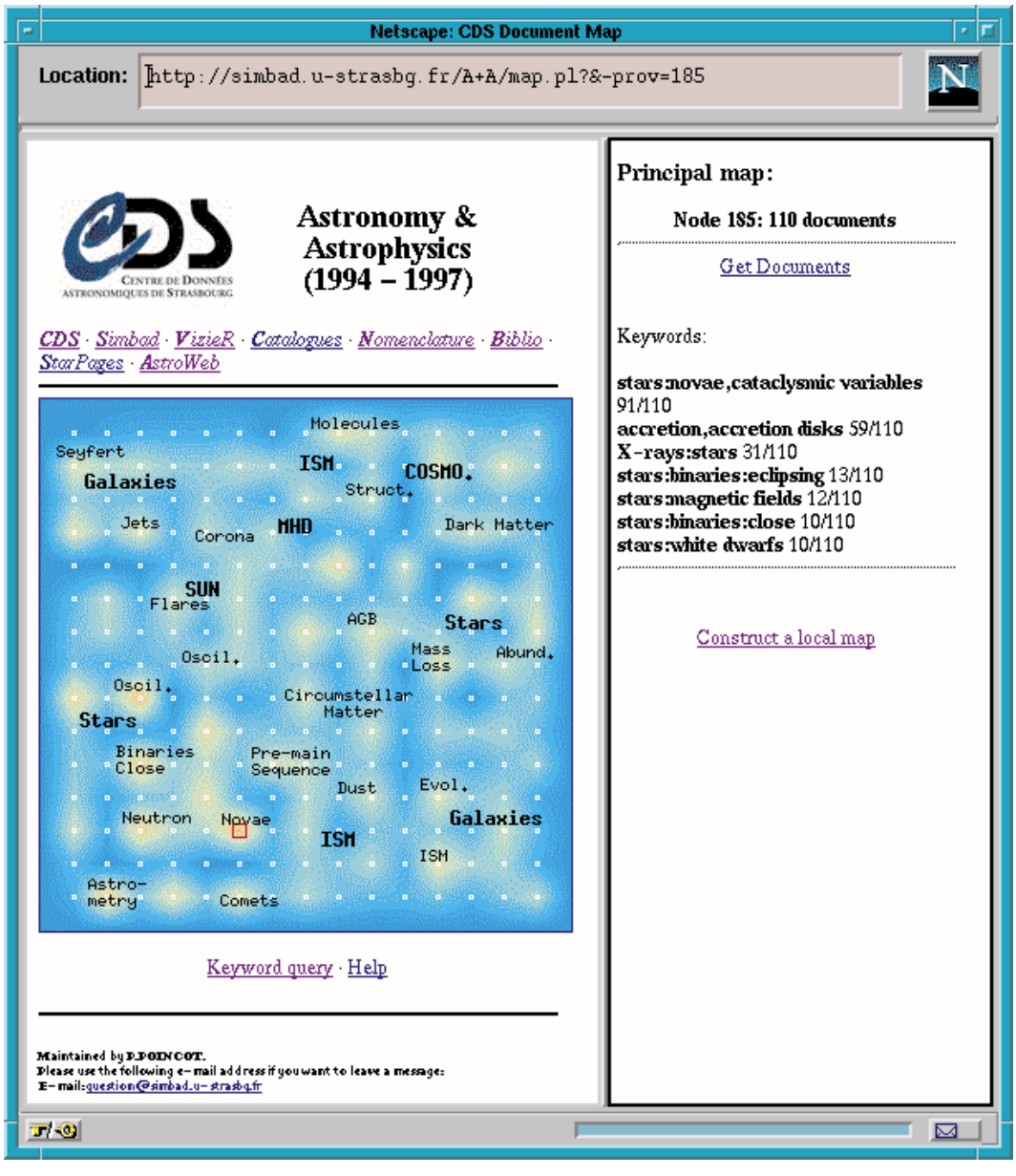

Fig. 5. Obtained by clicking in the Novae node on the opening page

6. Conclusions

Information browsing and retrieval, just as many other technology-based human activities, can be helped greatly by a good user interface and effective classification. The user interface developed in this work is an efficient and effective mechanism for browsing through bibliographic data, allowing access to the bibliographic and documentary databases of ADS and CDS.

A number of extensions of the present work are possible. An extensive assessment of other keyword systems is currently being undertaken. With continuing new additions to the bibliographic data, updates to the system will be similarly performed in a regular fashion. 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 access. These navigation tools are being made accessible through the *A&A* and VizieR access areas at the CDS. We expect that the reader, in using the Kohonen interface to the *A&A* data, will send us remarks and wishes, which we will take into consideration in future enhancements of our interface.

Acknowledgements. We are grateful for support for this work by D. Egret, Director, Strasbourg Observatory, and F. Genova, Director, CDS, Strasbourg Observatory. We also acknowledge with thanks early and enthusiastic comments by J. Lequeux on this work.

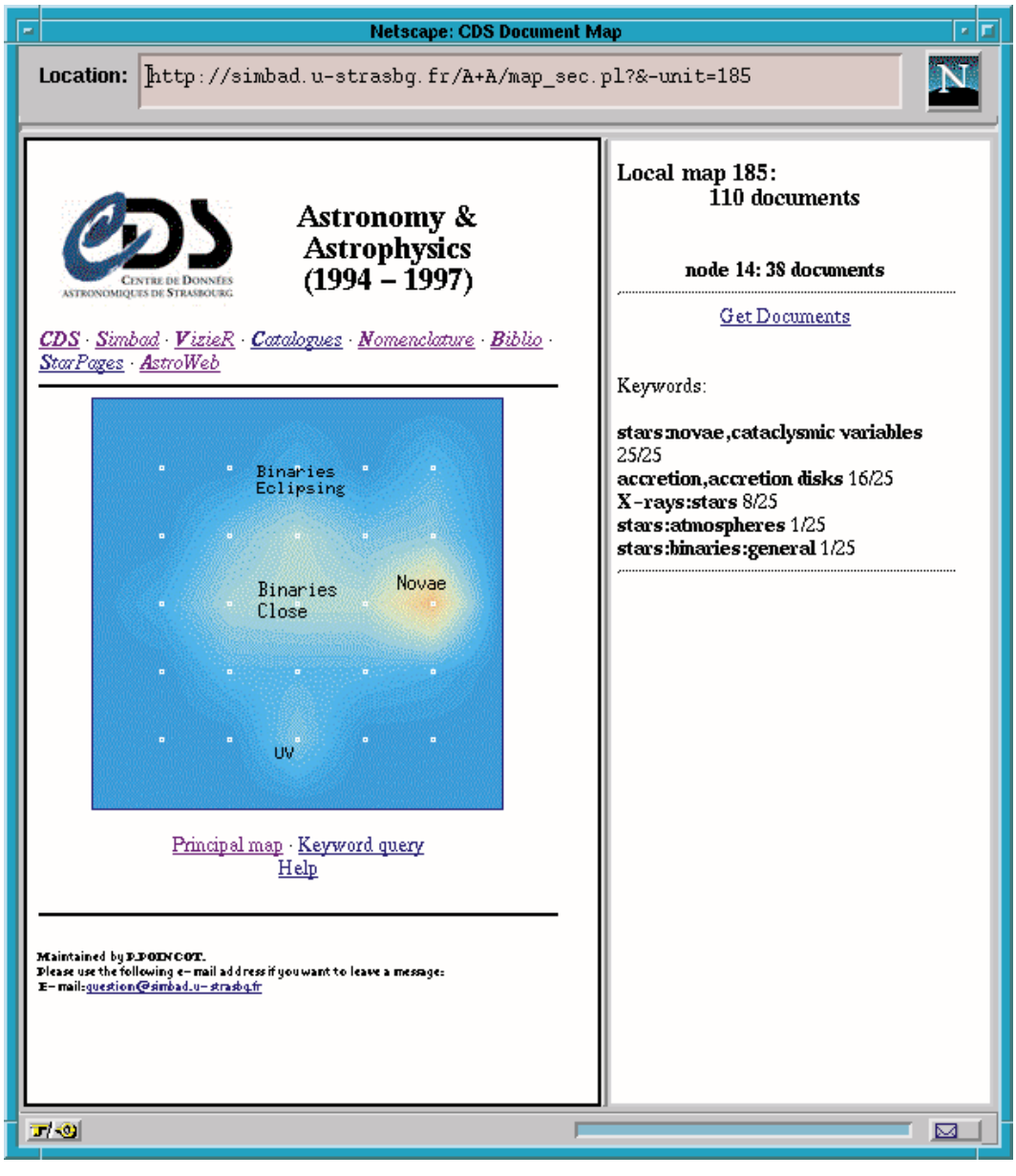

Fig. 6. The local map corresponding to the Novae node

References

Bhandarkar S.M., Koh J., Suk M., 1997, Neurocomputing 14, 241

Doyle L.B., 1961, J. ACM 8, 553

Egret D., Crézé M., Bonnarel F., 1995, in Astrophysics and Space Science Library 203, 163

Murtagh F., Hernández-Pajarez M., 1995, J. Classif. 12, 165

Murtagh F., Gopalan T.K., 1997, "Input data coding in multivariate data analysis: techniques and practice in correspondence analysis", in International Statistic Review (submitted)

Genova F., Bartlett J., Bienaimé O., et al., 1996, Vistas Astron. 40

Honkela T., Pulkki V., Kohonen T., 1995, Proc. ICANN-95, Int. Conf. on Artificial Neural Networks, II, 3

Kohonen T., 1995, "Self Organizing Maps". Springer-Verlag, Berlin

Lesteven S., Poinçot Ph., Murtagh F., 1995, Vistas Astron. 39, 187

Luttrell S.P., 1989, Pattern Recognition Lett. 10, 1

Pörner B., 1995, "VIRI - Visual Information Retrieval Interfaces - listing of information visualization systems", http://www-cui.darmstadt.gmd.de/visit/Activities/Viri/visual.html

Salton G., 1991, Sci 253, 974

WebSOM 1997, WEBSOM – Self-Organizing Map for Internet Exploration, various papers, interactive system, http://websom.hut.fi/

Zavrel J., 1996, Artif. Intel. Rev. 10, 477

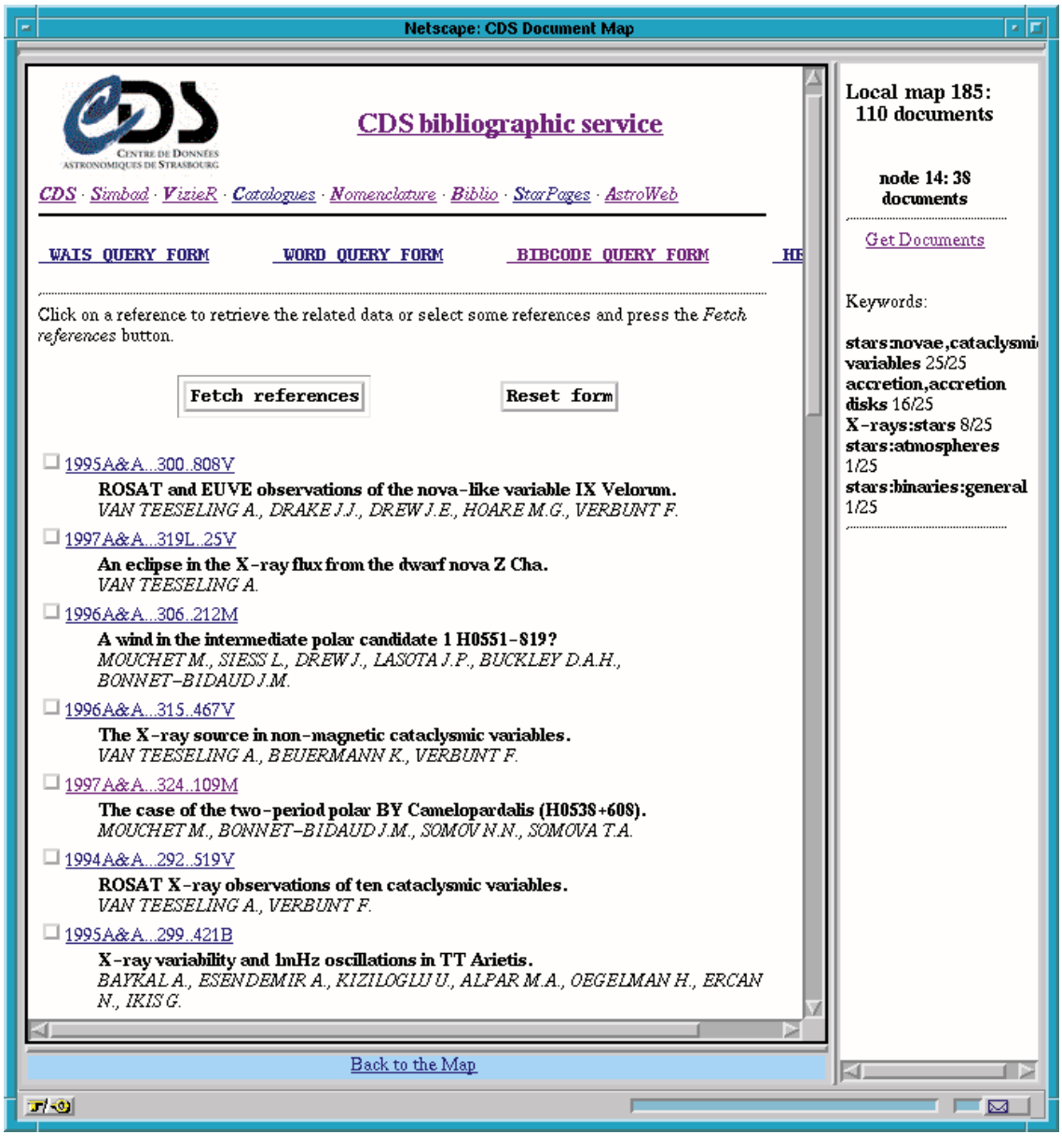

Fig. 7. List of articles associated with the secondary map. This is a CDS page proposing, inter alia, access to abstracts of articles