Spatial Organization Using Self-Organizing Neural Networks

Riccardo Rizzo, Marco Arrigo
Italian National Research Council-Institute for Educational and Training Technologies
Via Ugo La Malfa 153,
90146 Palermo,Italy
{rizzo,arrigo}@itdf.pa.cnr.it

ABSTRACT

Spatial hypertext systems use physical properties as color, dimensions, and position to represent relationships between documents. These systems allows the user to express a lot of different relationships between information but the structure should be build by hand by the user. This can be complex if a large number of information is involved. Self-organizing neural networks map can automatically generate a document map in which clusters of similar documents are organized. These maps can be used as a navigation tool "per se" or as a starting point for more complex spatial organizations. Systems based on SOM network can also automatically find the right map location for a new document, giving to the user a valuable help in information organization. In this paper the application of Self-Organizing Maps as a tool to develop information maps and spatial hypertext systems prototype is discussed and some applications are presented.

Keywords

Self-organizing networks, neural networks, visualization.

1. INTRODUCTION

In spatial hypertexts the user creates a two-dimensional space where documents and pieces of information are organized, and their relationships are expressed by using their relative location on a two dimensional space [5]. The freedom in building this information space allows the user to express a lot of different relationships using a "constructive ambiguity" [8] that is one of the major advantages of spatial hypertexts. In this paper spatial hypertexts refers to a set of systems that uses a spatial organization to represent a semantic structure.

An attempt to integrate statistical representation of the free text of a document and spatial organization was made in [2]. This approach brings to a system that is difficult to scale and to a visualization of the information space difficult to manage. In this paper we present another approach to automatic organization of information; this approach is based on a self-organizing neural network that automatically build a useful information map.

Self-Organizing Maps (SOM maps) [3] are artificial neural networks that can organize information or documents on a space using a two-dimensional array of neurons. The neurons can be considered as sensitive units capable to modify a set of parameters (their weights) in order to approximate an external input, during the learning stage. The mechanism of the learning stage will be explained in the following sections but in an informal way we can say that the when a document (represented by a set of suitable parameters) is submitted to the network the unit most sensitive (most similar to the document) usually called b.m.u. (best matching unit) is modified in order to become "more similar" (i.e. more sensitive) to the document itself. Different units can adapt their weights to different documents and on the surface of the map many specialized areas (set of units) will appear. This mechanism can be used for visualization and for organization of a set of document exploiting the same visual mechanism that is exploited in spatial hypertexts.

In the paper the application of SOM map to visual information organization and visualization is described. The results obtained show that the SOM is a viable help to automatically organize information on a two-dimensional space. In the next section the principle of the Self-Organized Map is explained, then the application prototypes are presented and some conclusions are drawn.

2. SELF-ORGANIZING NEURAL NETWORKS TO DEVELOP INFORMATION MAPS

Artificial neural networks (ANN) models are made up of a dense interconnection of simple non-linear computational elements corresponding to the biological neurons. Each connection is characterized by a variable weight that is adjusted, together with other parameters of the net, during the so-called "learning stage".

In Self-Organized Map the neurons are organized in a lattice, usually a one or two-dimensional array, which is placed in the input space and is spanned over the inputs distribution. Using a two-dimensional SOM network it is possible to obtain a map of input space where closeness between units or clusters in the map represents closeness of the input vectors. The SOM algorithm principle can be explained in an abstract system without reference to any biological structure. To each processing unit in the SOM lattice is associated a vector of weight of the same dimension of the input vectors. Using the weights of each processing unit as a set of coordinates the lattice can be positioned in the input space. During the learning stage the weights of the units change their position and "move" towards the input points as illustrated in figure 1. This "movement" becomes slower and at the end of the learning stage the network is "frozen" in the input space.

After the learning stage the inputs can be associated to the nearest network unit. When the map is visualized the inputs can be associated to each cell on the map. One or more cell that clearly contains similar documents can be considered as a cluster on the map. These clusters are generated during the learning phase without any other information. It is not necessary to supply to the network cluster prototypes or examples. The main applications of the SOM is the visualization of high-dimensional data in a two dimensional manner, and the creation of abstractions like in many clustering techniques. The characteristic that distinguishes the SOM net from the other classification algorithms is that not only similar inputs are associated to the same cell but also neighborhood cells contain similar documents. This property together with the easy visualization makes the SOM map a useful tool for visualization of large data sets.

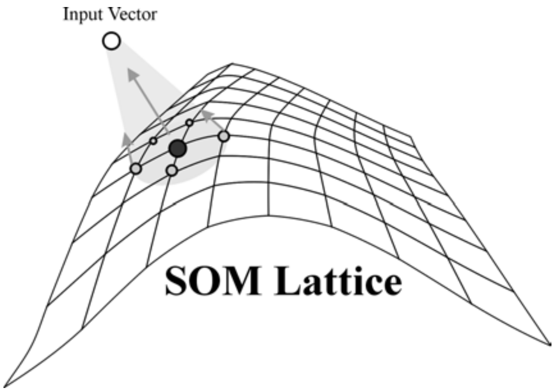


Figure1. Representation of the SOM learning phase.

Recently this network has been used to classify information and documents in "information maps". These are two-dimensional graphical representations in which all the documents in a document set are depicted. The documents are grouped in clusters that all concern the same topic, and clusters about similar topics are near each other on the map. To obtain these results documents must be represented by a vector using a Vector Space Representation as the TFIDF representation [9]. To build this document representation a set of meaningful keywords is chosen and each document is represented by a set of values each of them corresponding to the importance of the related keyword in the document.

Some studies indicate that the clustering results obtained using the SOM maps can be meaningful for the users. In particular in [4] was validated the proximity hypothesis for which related topics are clustered closely on the map.

In [7] the SOM was trained by using the nodes of a hypertext and the nodes in the same unit or in units connected by the rectangular lattice were considered linked each other. This organization was compared to the link structure imposed by the author and the number of the link in common was compared to the total number of links. The 64.5% of the link in the original hypertext was "covered" by the SOM network.

One of the most important results of this document organization is that the input space is also divided in many parts, one for each unit on the map, called Voronoi sets. Each unit of the SOM map groups together all the documents that are represented with a vector that belongs to that area. This area is a sensitive area for the neural unit. All the new documents which representing vector belongs to this area will be classified in the cluster of the neural unit. New documents will be automatically assigned to the cells of the map depending on their semantic characteristics, helping effectively the user in document organization. Using his previous knowledge about the document organization in the map space, the user can immediately recognize document's topics.

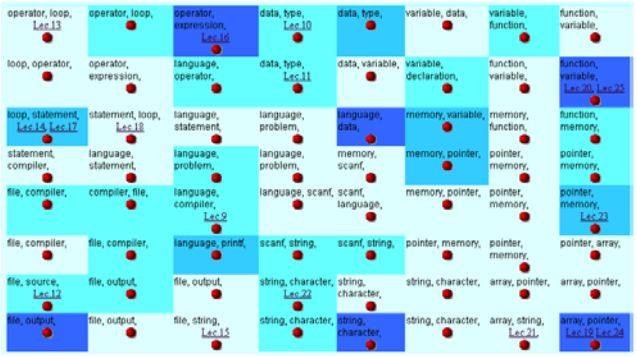
So each neural unit is sensitive to documents on a particular topic. In the following section we will see how to identify this topic.

3. ORGANIZATION OF LEARNING MATERIAL

In learning activities the user has to put in the same context and organize different learning materials from different sources: lesson handouts, book chapters, web documents, notes, and so on. Systems based on SOM build automatically a map that collect and organize pieces of knowledge, but they can make more.

One of the advantages of using this unsupervised learning neural network for clustering is that the user can know only a fraction of the material. For example a student can only remember the lessons and feel comfortable with the book chapters, but completely ignore the structure or the topics of other related learning materials. Using a trained SOM network the user can easily spot the known material having, in this way, an immediate idea of the spatial organization of the rest of the documents. The know material identify on the map areas that can be interesting for the user, and at the same time can give a clue of he topic of other unknown materials. The visual organization of the information becomes not only a way of navigate throughout the documents but also a way to focus the attention, finding information, and discarding what is not interesting.

One of the prototypes based on a SOM map is a system studied to support the navigation among different tutorial about the C language [1]. The map establishes a visual relationship among many different web documents that are picked up from different tutorials and web sites. The only contact point that they have is the cluster relationship established on the map.



operator, loop, Lec.13 ●	operator, loop, ●	operator, expression, Lec.16 ●	data, type, Lec.10 ●	data, type, ●	variable, data, ●	variable, function, ●	function, variable, ●
loop, operator, ●	operator, expression, ●	language, operator, ●	data, type, Lec.11 ●	data, variable, ●	variable, declaration, ●	function, variable, ●	function, variable, Lec.20, Lec.25 ●
loop, statement, Lec.14, Lec.17 ●	statement, loop, Lec.18 ●	language, statement, ●	language, problem, ●	language, data, ●	memory, variable, ●	memory, function, ●	function, memory, ●
statement, compiler, ●	language, statement, ●	language, problem, ●	language, problem, ●	memory, scanf, ●	memory, pointer, ●	pointer, memory, ●	pointer, memory, ●
file, compiler, ●	compiler, file, ●	language, compiler, Lec.9 ●	language, scanf, ●	scanf, language, ●	memory, pointer, ●	pointer, memory, ●	pointer, memory, Lec.23 ●
file, compiler, ●	file, compiler, ●	language, printf, ●	scanf, string, ●	scanf, string, ●	pointer, memory, ●	pointer, memory, ●	pointer, array, ●
file, source, Lec.12 ●	file, output, ●	file, output, ●	string, character, Lec.22 ●	string, character, ●	string, character, ●	array, pointer, ●	array, pointer, ●
file, output, ●	file, output, ●	file, string, Lec.15 ●	string, character, ●	string, character, ●	string, character, ●	array, string, Lec.21 ●	array, pointer, Lec.19 Lec.24 ●

Figure 2. The map of documents about C programming language.

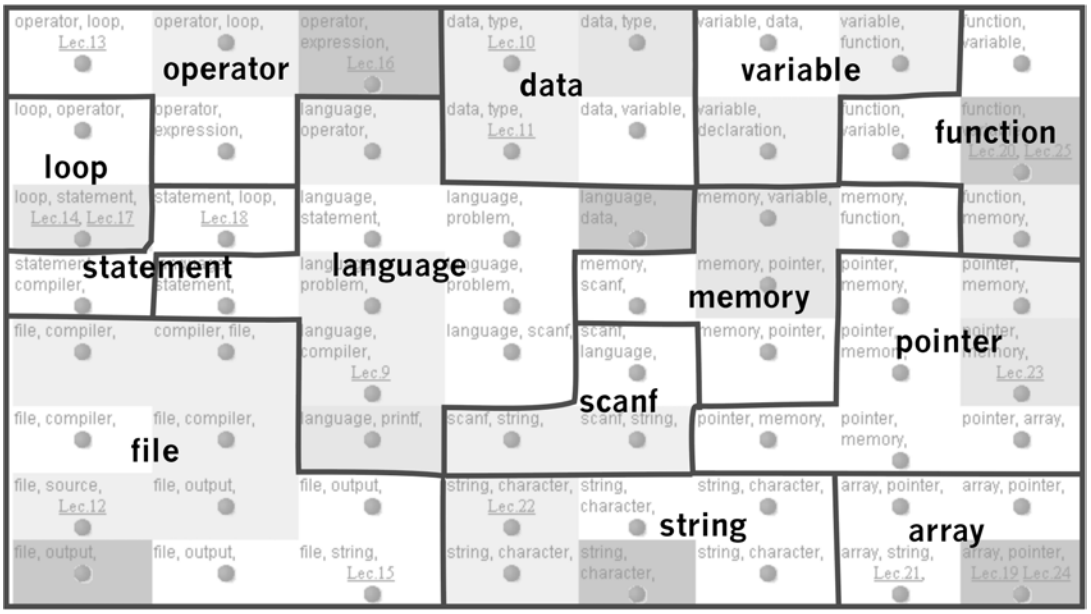
The system uses the lesson handouts as focus points for the students. These links related to the lessons allows the user to exploit the visual browsing of the resources. Students can look only at the areas that are interesting and can easily realize which is the topic related to each map cell. So they do not have to know the content or the organization of the different tutorials, they can simply look at the cells surrounding the lesson they are interested in.

The map, that is the main interface of the system, is shown in figure 2. The map is visualized as a HTML table where each cell is a neural unit. The map collects 590 Web pages collected from 6 on-line tutorials. Only 3 tutorials (114 web pages) were used for the training of the map, the others were added after the learning stage and automatically sorted by the neural network.

Inside the cells it is possible to see a red dot that the link to a list of document that are contained in the cell. This is necessary because some cells can contain more than 40 documents and a long list of documents inside the HTML table cell can be difficult to visualize. The quantity of documents inside each cell is indicated by the cell color (darker color indicates more documents).

Each cell of the map is labeled using a set of keywords (two in this case) that are calculated by the system. The two keywords are related to the two biggest weight values for that neural unit. These keywords give another help to identify the topic areas on the map and also give an idea of which is the sensitive area of he neural unit.

It is also possible to observe as the topics (i.e. the sensitive areas) gradually shift from one to another. For example on the bottom right of the figure 3 it is possible to notice that array, pointers and strings are closed together and as the two keywords of each cell change from a cell to another. In figure 3 it is also shown the set of the different topic area that can be recognized on he map. These areas are extracted looking at the first keyword of each single cell and grouping together the cells that have the same first keyword. Due to the "smooth" changing on the topic described before cells that share the same topic are near each other on the map. This map is used as a background image for another application that is presented in the next sections.

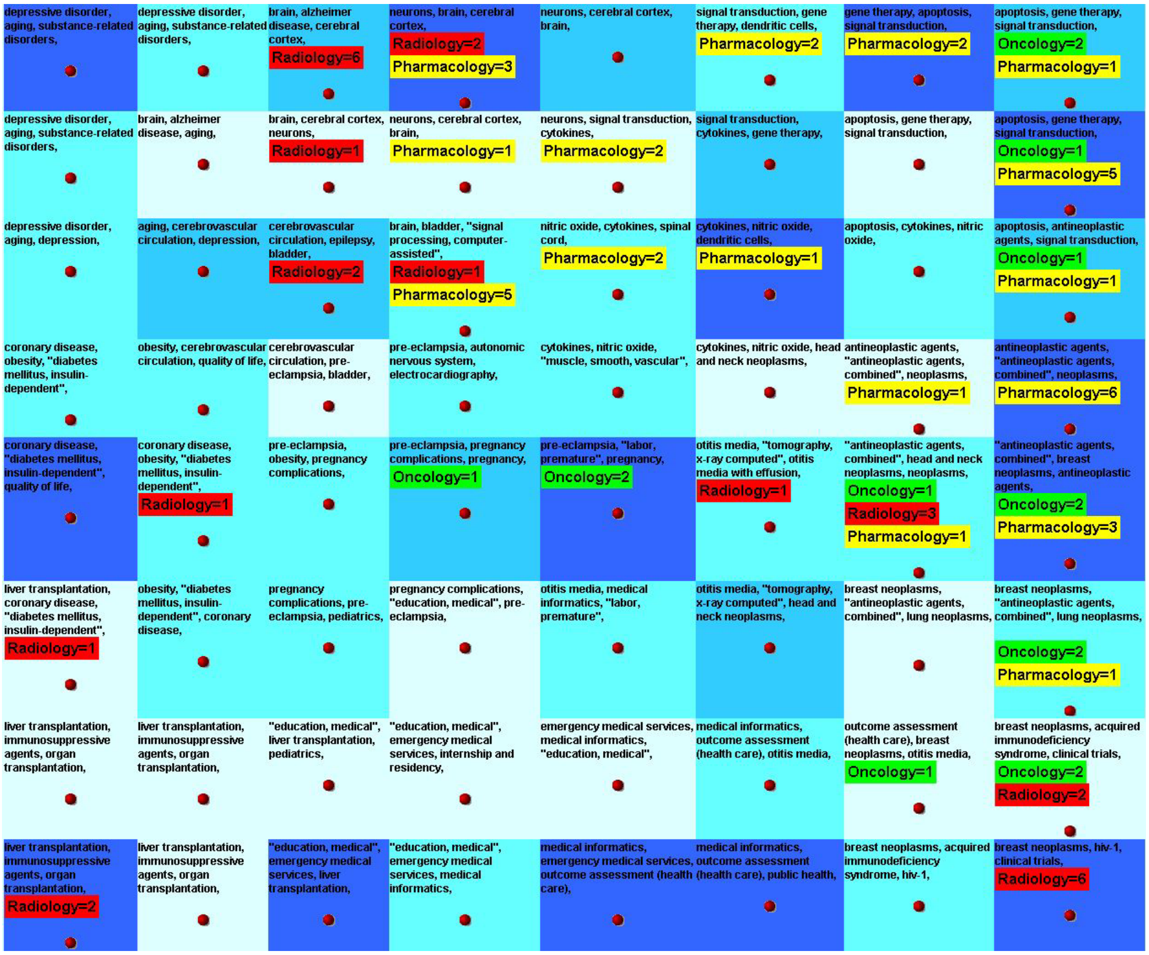


operator, loop, Lec.13	operator, loop,	operator, expression, Lec.16	data, type, Lec.10	data, type,	variable, data,	variable, function,	function, variable,
loop, operator,	operator, expression,	language, operator,	data, type, Lec.11	data, variable,	variable, declaration,	function, variable,	function, Lec.20, Lec.25
loop, statement, Lec.14, Lec.17	statement, loop, Lec.18	language, statement,	language, problem,	language, data,	memory, variable,	memory, function,	function, memory,
statement, compiler,	statement,	language, problem,	language, problem,	memory, scanf,	memory, pointer,	pointer, memory,	pointer, memory,
file, compiler,	compiler, file,	language, compiler, Lec.9	language, scanf,	scanf, language,	memory, pointer,	pointer, memory,	pointer, memory, Lec.23
file, compiler,	file, compiler,	language, printf,	scanf, string,	scanf, string,	pointer, memory,	pointer, memory,	pointer, array,
file, source, Lec.12	file, output,	file, output,	string, character, Lec.22	string, character,	string, character,	array, pointer,	array, pointer,
file, output,	file, output,	file, string, Lec.15	string, character,	string, character,	string, character,	array, string, Lec.21,	array, pointer, Lec.19 Lec.24

Figure 3. The background image that highlight the clusters of documents on the map.

4. VISUALIZATION OF DOCUMENTS INFORMATION

In the application presented in the previous sections the documents are represented using the free text in order to highlight their semantic similarity, but other information can be attached to a piece of knowledge and represented using a meta-data framework. Structured information, when not used for document representation, can be highlighted on the map in order to discover other interesting patterns. This property was exploited in a prototype that organizes 1353 personal records of medical researchers working in different topic areas and in different laboratories and institutions. The aim of the work was to highlight possible collaboration or information exchanges among different institutions or laboratories. Each personal record contains a unique id, the affiliation and a set of MeSH keywords (Medical Subject Headings) that characterize the working area or interest area for each researcher. This set of keywords was considered as a free text and used to build a representation for each researcher, and then the trained SOM map was used to visualize the affiliation or the id of the researchers. The obtained SOM represents a map of research topics; on this map it is possible to visualize the coverage of each institution, or the interest of single researcher. In figure 4 the map and the coverage of three laboratories is highlighted. The overlapping and the contact points between the three departments (the department of radiology, the department of radiation oncology and of pharmacy) can be easily seen and the common interests are highlighted. The number near the name of department indicates the number of people that work on that topic.



depressive disorder, aging, substance-related disorders,	depressive disorder, aging, substance-related disorders,	brain, alzheimer disease, cerebral cortex, Radiology=6	neurons, brain, cerebral cortex, Radiology=2 Pharmacology=3	neurons, cerebral cortex, brain,	signal transduction, gene therapy, dendritic cells, Pharmacology=2	gene therapy, apoptosis, signal transduction, Pharmacology=2	apoptosis, gene therapy, signal transduction, Oncology=2 Pharmacology=1
depressive disorder, aging, substance-related disorders,	brain, alzheimer disease, aging,	brain, cerebral cortex, neurons, Radiology=1	neurons, cerebral cortex, brain, Pharmacology=1	neurons, signal transduction, cytokines, Pharmacology=2	signal transduction, cytokines, gene therapy,	apoptosis, gene therapy, signal transduction,	apoptosis, gene therapy, signal transduction, Oncology=1 Pharmacology=5
depressive disorder, aging, depression,	aging, cerebrovascular circulation, depression,	cerebrovascular circulation, epilepsy, bladder, Radiology=2	brain, bladder, "signal processing, computer-assisted", Radiology=1 Pharmacology=5	nitric oxide, cytokines, spinal cord, Pharmacology=2	cytokines, nitric oxide, dendritic cells, Pharmacology=1	apoptosis, cytokines, nitric oxide,	apoptosis, antineoplastic agents, signal transduction, Oncology=1 Pharmacology=1
coronary disease, obesity, "diabetes mellitus, insulin-dependent",	obesity, cerebrovascular circulation, quality of life,	cerebrovascular circulation, pre-eclampsia, bladder,	pre-eclampsia, autonomic nervous system, electrocardiography,	cytokines, nitric oxide, "muscle, smooth, vascular",	cytokines, nitric oxide, head and neck neoplasms,	antineoplastic agents, "antineoplastic agents, combined", neoplasms, Pharmacology=1	antineoplastic agents, "antineoplastic agents, combined", neoplasms, Pharmacology=6
coronary disease, "diabetes mellitus, insulin-dependent", quality of life,	coronary disease, obesity, "diabetes mellitus, insulin-dependent", Radiology=1	pre-eclampsia, obesity, pregnancy complications,	pre-eclampsia, pregnancy complications, pregnancy, Oncology=1	pre-eclampsia, "labor, premature", pregnancy, Oncology=2	otitis media, "tomography, x-ray computed", otitis media with effusion, Radiology=1	"antineoplastic agents, combined", head and neck neoplasms, neoplasms, Oncology=1 Radiology=3 Pharmacology=1	"antineoplastic agents, combined", breast neoplasms, antineoplastic agents, Oncology=2 Pharmacology=3
liver transplantation, coronary disease, "diabetes mellitus, insulin-dependent", Radiology=1	obesity, "diabetes mellitus, insulin-dependent", coronary disease,	pregnancy complications, pre-eclampsia, pediatrics,	pregnancy complications, "education, medical", pre-eclampsia,	otitis media, medical informatics, "labor, premature",	otitis media, "tomography, x-ray computed", head and neck neoplasms,	breast neoplasms, "antineoplastic agents, combined", lung neoplasms,	breast neoplasms, "antineoplastic agents, combined", lung neoplasms, Oncology=2 Pharmacology=1
liver transplantation, immunosuppressive agents, organ transplantation,	liver transplantation, immunosuppressive agents, organ transplantation,	"education, medical", liver transplantation, pediatrics,	"education, medical", emergency medical services, internship and residency,	emergency medical services, medical informatics, "education, medical",	medical informatics, outcome assessment (health care), otitis media,	outcome assessment (health care), breast neoplasms, otitis media, Oncology=1	breast neoplasms, acquired immunodeficiency syndrome, clinical trials, Oncology=2 Radiology=2
liver transplantation, immunosuppressive agents, organ transplantation, Radiology=2	liver transplantation, immunosuppressive agents, organ transplantation,	"education, medical", emergency medical services, liver transplantation,	"education, medical", emergency medical services, medical informatics,	medical informatics, emergency medical services, outcome assessment (health care),	medical informatics, outcome assessment (health care), public health,	breast neoplasms, acquired immunodeficiency syndrome, hiv-1,	breast neoplasms, hiv-1, clinical trials, Radiology=6

Figure 4. The distribution of the research interest of three medical research laboratories.

5. FUTURE WORKS

The third prototype is in early stage of development and is an attempt to overcome some of the problems related to the SOM approach to the spatial organization of information.

The organization of the information space obtained by using the SOM is something that is not controlled by the user. The user can add documents on the map and use the neural network algorithm to put them in the right place, but it cannot move or change the position of the cluster labels. To this purpose another prototype was developed using the same documents and the same neural network. In this prototype the user can change the position of the clusters that are floating over a bitmap image that reproduces the clusters organization of the map, as shown in figure 5. This allows the user to change the visual organization of the cluster according to his preferences and overriding the neural network organization. Moreover the user is able to add new documents in any other position of the map. The prototype is based on a client-server architecture and is developed in Java. The system is in early phase of development and we plan to add many of the features discussed in the present paper, as the visualization feature in figure 4. We also plan to give to the user the freedom to change the shape and the color of the icons that represent the document clusters. We are also studying a way to have a "true" interaction to the neural network: in fact the changes in the position of the label of the map do not affect the "knowledge" of the map. The user can move the labels but the neural units remain fixed in the network lattice. So that the network is not aware of the changes made by the user. To obtain a complete interaction to the neural network a more complex mechanism is required that involves deeply the neural network model. In fact the position of the neural units and the shape of the lattice are part of the knowledge of the neural network, and they are the result of the learning phase over the document set.

6. CONCLUSIONS

The Self-Organizing Maps are effective tools to develop two dimensional information spaces. These information maps can be used to organize information and visualize their relationships as a spatial distribution. These results present SOM neural network as a tool to develop spatial hypertext and to help the user to manage their information space. At the moment the application is more a "batch" approach, where the network learn and present the result the user. An "interactive" approach should be studied is order to allow the user to guide the organization process, to modify the information space created or to make the neural network "aware" of the changes made by the user.

It should be noted that the SOM map is not the only self-organizing neural network that can be used to organize information on a lattice as it was shown in [6].

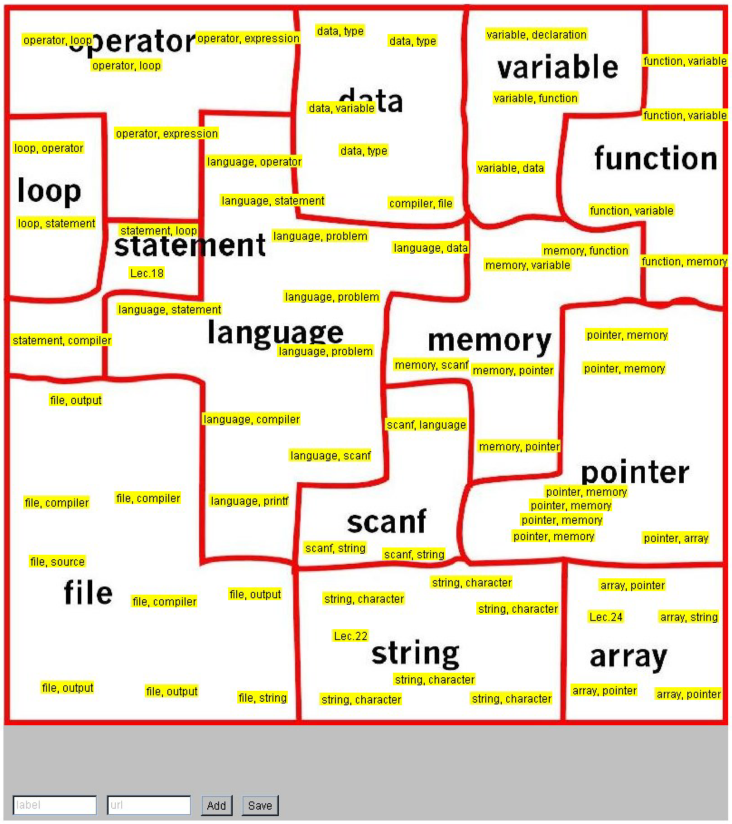


Figure 5. The spatial hypertext prototype in development.

7. REFERENCES

[1] Brusilowsky P., and Rizzo R., Map-Based Horizontal Navigation in Educational Hypertext, accepted at ACM Conference on Hypertext and hypermedia (Hypertext'02), in press.

[2] Chen, C. and Czerwinsky, M., From Latent Semantic to Spatial Hypertext – An Integrated Approach. In: Grønbæk, K., Mylonas, E. and Shipman III, F. M. (eds.) Proc. of Ninth ACM International Hypertext Conference (Hypertext'98), Pittsburgh, USA, ACM Press (1998) 77-86

[3] Kohonen, T., Self-Organizing Maps, Springer Verlag, Berlin, 1995.

[4] Lin, C., Chen, H., and Nunamaker, J. F. Verifying the Proximity Hypothesis for Self-Organizing Maps. In: Proc. of The 32nd Hawaii International Conference on System Sciences (1999)

[5] Marshall, C. C. and Shipman III, F. M. Spatial hypertext: Designing for change. Communications of the ACM **38**, 8 (1995)

[6] Rizzo R., Self Organizing Networks to Map Information Space in Hypertext Development, Proc. Of the International ICSC/IFAC Symposium on Neural Computation NC'98, September 23-25, 1998, Vienna, Austria.

[7] Rizzo, R., Allegra, M., and Fulantelli, G. Hypertext-like structures through a SOM network. In: Proc. of Tenth ACM Conference on Hypertext and hypermedia (Hypertext'99), Darmstadt, Germany, ACM Press (1999)

[8] Shipman III, F. M. and Marshall, C. C. Spatial hypertext: an alternative to navigational and semantic links. ACM Computing Surveys **31**, 4 (1999)

[9] van Rijsbergen C.J.. Information Retrieval. Butterworths, London, 1979.