Inference Bayesian Network for Multi-topographic neural network communication: a case study in documentary data

Shadi Al Shehabi, Jean-Charles Lamirel

Inference Bayesian Network for Multi-topographic neural network communication: a case study in documentary data

Shadi Al Shehabi
LORIA
Campus Scientifique, BP 239
54506 Vandoeuvre-lès-Nancy Cedex
France
Shadi.Al-Shehabi@loria.fr

Jean-Charles Lamirel
LORIA
Campus Scientifique, BP 239
54506 Vandoeuvre-lès-Nancy Cedex
France
Jean-Charles.Lamirel@loria.fr

Abstract

In this paper we present an original approach consisting in assimilating the behavior of the MultiSOM model, whose core model represents a significant extension of the classical Kohonen SOM model, to the one model of a Bayesian inference network. This approach is used both for validating the MultiSOM inter-map communication principles and for enhancing the accuracy of the probabilistic correlation computation mode that is already provided by the model In a complementary way, our approach also led us to prove that a neural multi-map model provided with unsupervised learning might well behave as a Bayesian inference network in which the estimation of posterior probabilities becomes a simple process only using prior similarity measures.

1. Introduction

This article deals with a particular neural network mapping innovation, which is called MultiSOM. The MultiSOM model was firstly introduced for the information retrieval purposes and it has been soon successfully tested for complex multimedia retrieval [11] and web mining tasks [1] [12]. The MultiSOM model is the multi-map extension of the Kohonen self-organizing map (SOM) model. A SOM map is constituted by a grid of nodes. The SOM learning process is an unsupervised classification process that is performed in such a way that related topics which are extracted from the dataset to be analyzed are mapped on neighboring nodes on the map. The MultiSOM model introduces the concepts of viewpoints and dynamics into the information analysis concept with its multi-maps displays and its inter-map communication mechanism. The dynamic information exchange between maps can be exploited by an analyst in order to perform cooperative deduction between several different analyses that have been performed on the same data.

In documentary database analysis the MultiSOM's viewpoint building principle consists in separating the description space of the documents into different subspaces corresponding to different keyword subsets. The viewpoints can fit into the structure of the document when they correspond to different index vocabulary subsets associated to the different document sub-fields. Moreover, specific viewpoints may also be associated to specific reference fields like document bibliographies or document hyperlinks. As soon as each viewpoint is associated to a peculiar set of maps, the MultiSOM's inter-map communication mechanism enables an analyst to highlight semantic relationships between different topics belonging to different viewpoints. As an example, in patents analysis the model permits to detect association existing between the components used in patents, the advantages of the solutions described in these later, their domain of use and the patentees associated to the patents [13].

The paper shows that the Bayesian inference network to which the MultiSOM model can be assimilated is constituted of three layers: the first layer contains the nodes of the source map the second layer contains the intermediary nodes (i.e. data nodes), and the third layer contains the nodes of the target map. The standard Bayesian inference network propagation algorithm is used to compute the posterior probabilities of target map's cluster nodes which inherited of the activity transmitted by its associated data nodes.

2. Self-Organizing Map (SOM)

The basic principle of the SOM is that our knowledge organization at higher levels is created during learning by algorithms that promote self-organization in a spatial order (see [5], [6], [7], [8], [9] and [10]). Thus, the architecture form of the SOM network is based on the understanding that the representation of data features might assume the form of a self-organizing feature map that is geometrically organized as a grid or lattice. In the pure form, the SOM defines an "elastic net" of points (parameter, reference, or codebook vectors) that are fitted to the input data space to approximate its density function in an ordered way. The algorithm takes thus a set of N-dimensional objects as input and maps them onto nodes of a two-dimensional grid, resulting in an orderly feature map [7]. A layer of two-dimensional array of competitive output nodes is used to form the

0-7803-8482-2/04/$20.00 ©2004 IEEE.

feature map. The lattice type of array can be defined to be square, rectangular, hexagonal, or even irregular. Every input is connected to every output node via a variable connection weight. It is the self-organizing property. The SOM belongs to the category of the unsupervised competitive learning networks [9] [14]. It is called competitive learning because there is a set of nodes that compete with one another to become active. To this category belongs also the adaptive resonance theory (ART) model of Grossberg and Carpenter, as well as the self-organizing multiple maps discussed in this paper. In the SOM, the competitive learning means also that a number of nodes is comparing the same input data with their internal parameters, and the node with the best match (say, "winner") is then tuning itself to that input, in addition the best matching node activates its topographical neighbours in the network to take part in tuning to the same input. The more a node is distant from the winning node the weaker is the learning. It is also called unsupervised learning because no information concerning the correct classes is provided to the network during its training. Like any unsupervised classing method, the SOM can be used to find classes in the input data, and to identify an unknown data vector with one of the classes. Moreover, the SOM represents the results of its classing process in an ordered two-dimensional space (R^2). A mapping from a high-dimensional data space R^n onto a two dimensional lattice of nodes is thus defined. Such a mapping can effectively be used to visualize metric ordering relations of input data. As Kohonen [7] says: "The main applications of the SOM are in the visualization of complex data in a two dimensional display, and creation of abstractions like in many classing techniques."

The SOM algorithm is presented in details in ([6], [7],[14]). It consists of two basic procedures: (1) selecting a winning node and (2) updating weights of the winning node and its neighbouring nodes. This preliminary learning phase is not straightforward process [7]. It necessitates several different learning steps, single map evaluations, and comparisons between a lot of generated maps in order to find at least a reliable map, at most an optimal one [14].

a - Winning node selection:

Let $x(t) = \{ x_1(t), x_2(t), \ldots, x_N(t) \}$ be the input vector selected at time t, and $W_k(t) = \{ W_{k1}(t), W_{k2}(t), \ldots, W_{kN}(t) \}$ the weights of the codebook vector associated to the node k at time t. The smallest of the Euclidean distances $\| x(t) - W_k(t) \|$ can be used to define the winning node s:

$$\| x(t) - W_s(t) \| = \min \| x(t) - W_k(t) \|$$

b - Unsupervised learning and definition of the neighbourhood:

After the winning node s thus selected, the weights of its codebook vector and the weights of the codebook vectors of the nodes in a defined neighbourhood (for example all nodes within a square or a cycle around the winning node) are adjusted so that similar input patterns are more likely to select these nodes again. This is achieved through the following computation:

$$W_{ki}(t+1) = W_{ki}(t) + \alpha(t) \times h(t) \times [X_i(t) - W_{ki}(t)],$$
for $1 \leq i \leq N$

where $\alpha(t)$ is a gain term ($0 \leq \alpha(t) \leq 1$) that decreases in time and converges to 0, and $h(t)$ is the neighbourhood function.

Once the SOM algorithm is achieved, the data can be affected to the nodes of the map. For each input data vector, the winning node is selected according to the algorithm first step presented above, and the data are affected to this selected node.

In the quantitative studies of science, the Kohonen self-organizing maps have been successfully used for mapping scientific journal networks, and also author co-citation data. Maps have been also successfully used for several other applications in the general area of data analysis like for classifying meeting output, for classing socio-economic data and for documentary database contents mapping and browsing [15] [16].

3. The MultiSOM model

The communication between self-organizing maps has been firstly introduced in the context information retrieval for analyzing the relevance user's queries regarding to the documentary database contents [14]. It represents a major amelioration of the basic Kohonen SOM model. From a practical point of view, the MultiSOM model introduces the use of *viewpoints* in the information analysis. Each different viewpoint is achieved in the form of a map along with its potential generalizations. Each single map used in the MultiSOM model is itself a spatial order in which the information is both represented into nodes (classes or topics) and spatial areas (group of classes or macro-topics) [14] [15]. The inter-map communication mechanism enables a user to highlight semantic relationships between different topics belonging to different viewpoints.

One of the main advantages of the MultiSOM model as compared to classical models that has been both highlighted by human expert and by objective evaluation [12] is that the original multiple viewpoints classification tends to reduce the noise which is inevitably generated in an overall classification approach, like SOM, while thoroughly increasing the flexibility and the granularity of the analyses.

3.1. The viewpoint notion

The viewpoint building principle consists in separating the description space of the documents into different subspaces corresponding to different keyword subsets.

The set of V all possible viewpoints issued from the description space D of a document set can be defined:

$$V = \{v_1, v_2, \ldots, v_n\},\ v_i \in P(D), \text{ with } \bigcup_{i=1}^{n} v_i = D$$

where each v_i represents a viewpoint and P(D) represents the set of the parts of the description space of the documents D; the union of the different viewpoints constitutes the description space of the documents.

The viewpoint subsets issued from V may be overlapping ones. They also fit into the structure of the document when they correspond to different index vocabulary subsets associated to the different document sub-fields. In the context of a documentary database, specific viewpoints may be associated to specific reference fields like "indexer keywords", "title keywords", or "author" field. Complementary viewpoints may be also extracted from the overall document description space. The notion of viewpoint is more general than the one of document field. It is always possible to find a viewpoint that represents the description space used in a document field.

3.2. Inter-map communication mechanism

In MultiSOM, this communication is based on the use of the data that have been projected onto the maps as intermediary nodes or activity transmitters between maps. The intercommunication process between maps operates in three successive steps. Figure 1 shows graphically the three steps of this intercommunication mechanism.

At the step 1, the original activity is directly set up by the user on the node or on the logical areas of a source map through decisions represented by different scalable modalities (full acceptance, moderated acceptance, moderated rejection, full rejection) directly associated to nodes activity levels. This procedure can be interpreted as the user's choices to highlight (positively or negatively) different topics representing his centers of interest relatively to the viewpoint associated to the source map. The original activity could also be indirectly set up by the projection of an user's query on the nodes of a source map. The effect of this process will then be to highlight the topics that are more or less related to that query. Therefore, the activity of each map node is set up to the value of the cosine correlation measure [18] (see equation (1)) between the node vector and the query vector. The activity transmission to target maps is based itself on two elementary steps: a first transmission step from the activated source map to its associated document nodes (down activation), and a second transmission step from the activated document nodes to the target map (up reactivation). The activity of a class i of the target map T derived from the activity of a source map S is computed according to the following formula:

$$A_i^T = f_{n \in i}(g(A_n)),\ A_n = g(A_{j_n}^S)$$

where n represents a node associated to a data, j_n its associated class on the source map, f is a function implementing the semantic correlation computation described below, g is a bias function also described in the next paragraphs.

This activity transmission can be considered as a process of evaluation of the semantic correlation existing between topics of a source viewpoint (source map), and topics belonging to several other viewpoints (target maps).

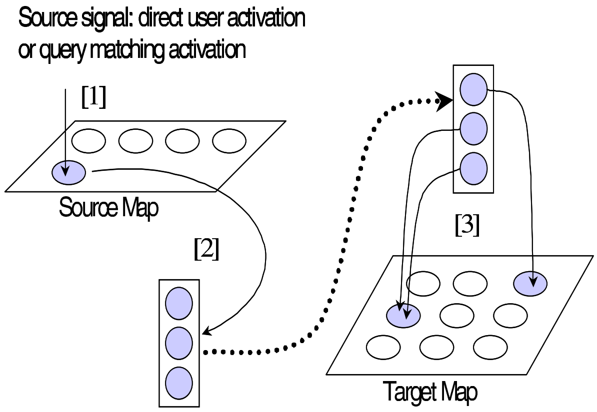


Figure 1: Inter-Map Communication Mechanism. This figure represents the main steps of the inter-map communication mechanism. (1) The activity is set up directly by the user or by a query formulation on one or several nodes of one or several source map. (2) The activity is transmitted to the data nodes associated to the activated class nodes of the source map. (3) The activity is transmitted through the data nodes to other maps to which these data are associated. Positive as well as negative activity could be managed in the same communication process. Note that the data are in this case indexed documents.

The parameters of the intercommunication procedure that are proposed to the decision of the analyst are the two modes of computing the semantic correlation, *f*, a possibilistic mode or a probabilistic mode, and the use of the bias function *g*.

In the *possibilistic computing of the semantic correlation* each class inherited of the activity transmitted by its most activated associated data. The *f* function can be given as:

$$f = \underset{n \in i}{Max}\left(A_n^+\right) + \underset{n \in i}{Max}\left(A_n^-\right)$$

where A+ represents a positive activity value (positive choice), and A^- a negative activity value (negative choice). This approach helps the user to detect weak semantic correlation (weak signals) existing between topics belonging to different viewpoints. For possibilistic theory, see [2].

In the *probabilistic computation of the semantic correlation*: each class inherited of the average activity transmitted by its associated data, either they are activated or not. The f function associated to the probabilistic mode can then be given as:

$$f = \frac{1}{|i|}\sum_{n \in i} A_n$$

where |i| represents the number of data associated to the class i. The probabilistic computation gives a more reliable measure of the strength of the semantic correlations, and may be then used to differentiate between strong and weak matching.

The role of the bias function g, which can be optionally used, is to modulate the activity transmission from a class to a data (down activation), and afterwards from a data to a class (top activation), considering the belonging degree of a data to a class as an attenuation factor for that transmission. The belonging degree of a document D_i to a class S_i is computed thanks to the cosine correlation measure [18]:

$$Sim(D_i, S_i) = \frac{D_i \bullet S_i}{\|D_i\|\|S_i\|} \qquad (1)$$

where $\|D_i\|$ represents the norm of the index vector associated to the document D_i, $\|S_i\|$ the norm of the codebook vector associated to the class S_i, and • represents the scalar product.

To perform in the best conditions, the inter-map communication process obviously necessitates that a significant part of the data should play that roles between the maps. This last condition could be easily verified if each vector used for the map generation indexes a significant part of the bibliographic database.

5. Inference Bayesian Network

A Bayesian network $G = (V,E)$ is a Directed Acyclic Graph (DAG), where the nodes in V represent the random variables [17] associated to the problem to be solved, and the arcs in E represent the dependence relationships among these variables. In such kind of graph, the knowledge is represented in two ways: (a) Qualitatively, showing the (in)dependencies between the variables, and (b) Quantitatively, by means of conditional probability distributions which shape the relationships. Thus, each variable $X_i \in V$ is provided with a family of conditional probability distributions $P(X_i \mid Pa(X_i))$, where $Pa(X_i)$ represents the parent set of the variable X_i in G [3] [4].

Bayesian networks can perform efficiently reasoning tasks: the independencies represented in the graph reduce changes in the state of knowledge to local computations. There are several algorithms that exploit this property to perform probabilistic inference (propagation), i.e., to compute the posterior probability for any variable given some evidence about the values of other variables in the graph.

6. Inference Bayesian Network model for Inter-map communication

The Bayesian inference network to which the MultiSOM model can be assimilated is constituted of three layers: the first layer contains M nodes of the source map S which may be activated by the analyst, the second layer contains N intermediary nodes D (i.e. data nodes) which are associated to the nodes of the source map, and the third layer contains M nodes of the target map T which may be activated through the intermediary nodes D that it shares with the source map.

Focusing on the structure of the network, the following guidelines have been considered to determine the topology of the graph:

1. The relationships between classes in source and destination maps only occur through the documents included in these classes.
2. Links joining classes and documents must be directed from source class nodes to document nodes and from document nodes to destination class nodes.

Thanks to the map construction process (see section 2), there is only one parent for each document node D_j such that the source class contain this document, then:

$$Pa(D_j) = S_i \in S \qquad (2)$$

The probability of a class S_i of the source map, considering evidence Q, can be itself estimated as:

$$P(S_i \mid Q) = \frac{1}{M} \qquad (3)$$

The parent set of any destination class node T_i, is represented by the set of document nodes that belong to T_i, i.e., $Pa(T_i) = \{D_j \in D \mid D_j \in T_i\}$. In the original MultiSOM model, there are not links between the document nodes and class nodes such that the source class nodes could represent the root nodes. Hence, this resulted in associating to the MultiSOM model a Bayesian network topology which is constituted of three layers.

The standard Bayesian inference network propagation algorithm is used to compute the posterior probabilities of target map's class nodes T_k which inherited of the activity (evidence Q) transmitted by its associated data nodes D. These computations can be carried out efficiently because of the specific Bayesian inference network topology that can be associated to the MultiSOM model. Hence, it is possible to compute the probability $P(act_m|T_k,Q)$ for an activity of modality act_m[1] on the class T_k which is inherited from activities generated on the source map. This computation is achieved as follows:

$$P(act_m|T_k,Q) = \frac{P(act_m \cap T_k|Q)}{P(T_k|Q)} \quad \text{(Bayes)} \qquad (4)$$

$$P(act_m \cap T_k|Q) = \sum_{D_j \in act_m \cap T_k} P(D_j|Q) \qquad (5)$$

$$P(T_k|Q) = \sum_{D_j \in T_k} P(D_j|Q) \qquad (6)$$

$$P(D_j|Q) = \sum_{i=1}^{M} P(D_j|S_i).P(S_i|Q) \qquad (7)$$

from (3) and (7), we obtain:

$$P(D_j|Q) = P(D_j|S_i).P(S_i|Q) \qquad (8)$$

from (5), (6) and (8), we derive:

$$P(act_m|T_k,Q) = \frac{\sum_{D_j \in act_m \cap T_k} P(D_j|S_i).P(S_i|Q)}{\sum_{D_j \in T_k} P(D_j|S_i).P(S_i|Q)} \qquad (9)$$

$$P(D_j|S_i) = \frac{P(D_j \cap S_i)}{P(S_i)} \quad \text{(Bayes)} \qquad (10)$$

At that step, as $P(D_j \cap S_i)$ can be considered as a measure of correlation between D_i and S_i, we make the hypothesis that it can be assimilated to cosine correlation measure $Sim(D_j,S_i)$ (see equation (1)). Hence, the cosine correlation measure ranges in the interval *[0,1]*. The more the two vectors associated to D_i and S_i are (positively) correlated, the nearest is this measure to 1. The less they are correlated, the nearest is this measure is to 0. Thus:

$$P(D_j \cap S_i) = Sim(D_j,S_i) \qquad (11)^2$$
$$0 \le Sim(D_j,S_i) \le 1 \ .$$

from the combination of (9) and (11), we derive:

$$P(act_m|T_k,Q) = \frac{\sum_{D_j \in act_m \cap T_k} Sim(D_j,S_i)}{\sum_{D_j \in T_k} Sim(D_j,S_i)} \qquad (12)$$

[1] [14] shows how each user's decision modality (full acceptance, moderated acceptance, moderated rejection, full rejection) can be associated with a numerical value.

[2] To our opinion, this hypothesis is more general and more accurate that the one which is achieved in [17]. In this latter approach, the cosine correlation measure is used to represent the posterior probability of a document relatively to a given query *P(document | query)*. Hence, cosine correlation measure is more suitable to express correlation than similarity in a probabilistic approach.

Finally, the probabilities of all the activities, corresponding to the different modalities, on the class T_k in the destination map can be represented as:

$$P(Act|T_k,Q) = \begin{pmatrix} P(act_1|T_k,Q) \\ P(act_2|T_k,Q) \\ .... \\ P(act_i|T_k,Q) \\ .... \\ P(act_r|T_k,Q) \end{pmatrix} \qquad (12)$$

The eq. (12) permits to select which modality has the best probability for the class T_k.

7. Experimental results

Our experiment consisted in comparing the performance of our two activity propagation models, i.e. the original probabilistic mode, provided with MultiSOM, and the Inference Bayesian Network model. For that purpose we constructed two differents maps. Our test database is a database of 1000 patents that has been used in one of our preceding experiments. For the viewpoint-oriented approach the structure of the patents has been parsed in order to extract two different subfields corresponding to two different viewpoints: Use and Advantages. As it is full text, the content of the extracted fields associated with the different viewpoints is parsed by a lexicographic analyser [19] in order to extract viewpoint specific indexes. For each specific viewpoint the resulting index set is weighted by means of an *IDF* weighting scheme [20] and a map of 10x10 neurons (classes) is finally generated.

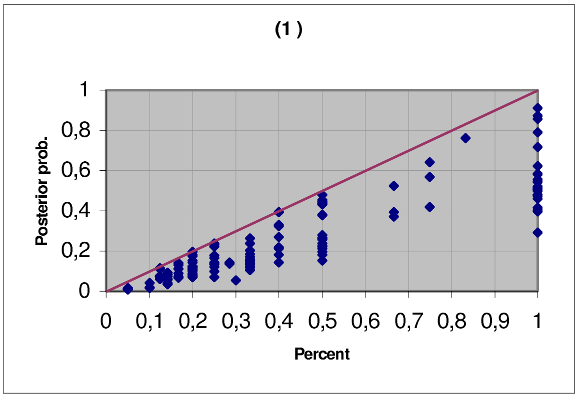
In our experiment the map associated to the Use viewpoint is considered as the source map as well as the map associated to the Advantage viewpoint is considered as the destination map. The testing process consists in launching separately on all the classes of the source map an activity corresponding to a full acceptance modality and then observing the transmitted activity to the classes of the destination map. The results of this process are presented in the figure 2.

Focusing on the results in the figure 2 we can conclude that the Bayesian model is more linear than the probabilistic mode. Hence, one can easily conclude by the observation of the figure 2 that the deviation of the probability for the same number of activated data in the case 1 is smaller than that in the case 2. This also means that in the case of the Bayesian model the results are more similar one to another for a given proportion of active documents and the proportion of active documents in a class of a destination map is more compliant with the probability of activity of this class. This latter phenomenon is especially obvious both for very low and very high probability values.

8. Conclusion

We have proposed a propagation mode based on a Bayesian inference network model for the MultiSOM model. We have shown that its results are more reliable than the probabilistic mode which was originally provided by the MultiSOM model. Moreover, it would be easy to prove that the Bayesian model is the only one that guaranties the coherence of the results in real world examples when various modalities will activate different documents of the same classes of destination map.

As soon is not limited to a specific subset of decision modalities our propagation model can already be applied to various kinds of model of communicating classification based on the notion of viewpoints. We nevertheless plan to render it still more general by extending it to a continuous user's decision model.



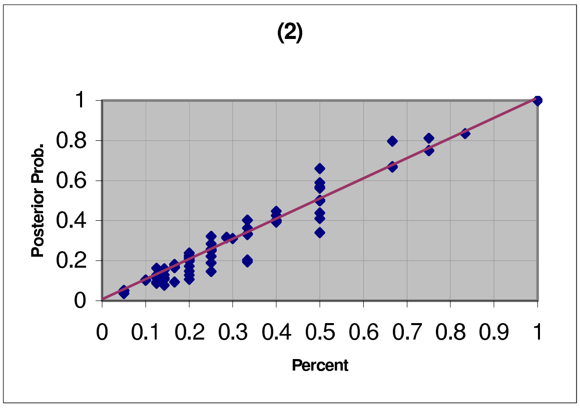


Figure 2: The figure illustrates the behaviour of the two models based on the relationship between the percentage of activated data in the classes of the destination map and the posterior probability measure generated by the propagation process: (1) the probabilistic mode, (2) inference Bayesian network model.

9. References

[1] IST-1999-20350, EICSTES project.

[2] D. Dubois, H. Prade, Théorie des possibilités, Masson, Paris, 1987.

[3] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete: A Layered Bayesian Network Model for Document Retrieval. ECIR 2002: 169-182.

[4] Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete: Building Bayesian Network-Based Information Retrieval Systems. DEXA Workshop 2000: 543-552.

[5] G. E. Hinton, «Connectionist Learning Procedures,» Artificial Intelligence, 40 (1989) p. 185-234.

[6] T. Kohonen, Self-Organisation and Associative Memory**,** Springer Verlag, Third edition, Berlin, 1984.

[7] T. Kohonen «The Self-Organizing Map,» Proceedings of the IEEE, 78 (1990) No 9, p. 1464-1480.

[8] T. Kohonen, «Self-Organizing Maps: Optimization Approaches,» in Artificial Neural Networks, T. Kohonen, K. Mäkisara, O. Simula, J. Kanges, Editors, Elsevier Science Publishers B.V, North Holland, Amsterdam, 1991, p. 981–990.

[9] T. Kohonen, «Things You Haven't Heard about the Self-Organizing Map,» IEEE International Conference on Neural Networks, San Francisco, Calif., March 28 – April 1, (1993) p. 1147-1156.

[10] T. Kohonen, Self-Organizing Maps. Springer Verlag, Berlin, 1997.

[11] J. C. Lamirel, « Using images for enhancing discovery tasks in a digital library context ». Proceeding of SPIE2001, San José, CA, January 2001.

[12] J.C. Lamirel, S. Al Shehabi, C. Francois, M. Hoffmann, « New classification quality estimators for analysis of documentary information: application to web mapping ». Proceedings of ISSI, Bejing, 2003.

[13] J.C. Lamirel, S. Al Shehabi, M. Hoffmann, C. Francois, « Intelligent patent analysis through the use of a neural network : experiment of multi-viewpoint analysis with the MultiSOM model ». Proceeding of ACL, Sapporo, Japan.

[14] J.C. Lamirel, Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif, Thèse de l'Université de Nancy 1 Henri Poincaré, 1995.

[15] X. Lin, D. Soergel, G. Marchionini, «A Self-Organizing Semantic Map for Information Retrieval,» in Proceedings of the 4th International SIGIR Conference on R&D in Information Retrieval, 13-16 October, Chicago, 1991, p. 262-269.

[16] X. Lin., « Map Displays for Information Retrieval, » JASIS, 48 (1) : 40-54, 1997.

[17] B.A. Ribeiro-Neto and R.R. Muntz. A belief network model for IR. In Proceedings of the 19 ACM--SIGIR Conference on Research and Development in Information Retrieval, pages 253--260. 1996.

[18] G. Salton, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971.

[19] O. Jouve, «Les nouvelles technologie de la recherche d'information», Séminaire Documentation, Paris, Otobre, 1999

[20] S. E. Roberston and K. Spark Jones, «Relevance Weighting of Search Terms», Journal of the American Society for Information Science, 27:129-146, 1976.