Visual research topic discovery with MULTI-SOM model of analysis

Xavier Polanco and Martial Hoffmann
Unité de Recherche et Innovation (URI)
Institut de l'Information Scientifique et Technique (INIST- CNRS)
2, allée du Parc de Brabois – 54514 Vandoeuvre-lès-Nancy – France
polanco@inist.fr, hoffmann@inist.fr

This article deals with a method of research topic discovery and visualization. Its implementation is called MULTI-SOM. The prefix MULTI indicates an extension of the Kohonen's Self-Organizing Map (SOM) in terms of multiple maps. This model is founded on the concept of viewpoint, and the implementation of the viewpoints into SOMs shaped in logical zones. The multiple maps are related by an original mechanism of communication. The system includes a mechanism of labelling the clusters, and also a mechanism of generalization. We apply MULTI-SOM to 832 bibliographical data about "Enterprise and Internet," following five different points of view.

Cet article expose une méthode pour la découverte et la visualisation de thèmes de recherche. Son implémentation est dénommée MULTI-SOM. Le préfixe MULTI signifie une extension des cartes auto-organisatrices de Kohonen (SOM) en termes de cartes multiples. Ce modèle est fondé sur le concept de point de vue, et la mise en œuvre des points de vue sous la forme de cartes auto-organisatrices. Les cartes multiples sont reliées entre elles par un mécanisme original de communication. Le système inclut la capacité d'étiqueter les classes, et également un mécanisme de généralisation. MULTI-SOM est ici appliqué sur un ensemble de 832 données bibliographiques, au sujet de l'« Entreprise et Internet », suivant cinq points de vue différents.

The implementation of the MULTI-SOM model that we present in this paper is the result of a research cooperation in the European project EICSTES (IST-1999-20350), between the research teams CORTEX-LORIA (UMR 7503) and URI-INIST (UPS 076), in the Fifth Framework Program on Research and Development of the European Union.

Keywords: Artificial Neural Networks, Clustering, Cartography, Visualization, Information Analysis, Self-Organizing Map, Information Society, Internet, Economics

1. INTRODUCTION

This article deals with a model of analysis of information and its technological implementation that we call MULTI-SOM. The model that we note by the prefix MULTI is implemented on the basis of Self-Organizing Maps (SOM) of Kohonen (1997). The SOM algorithm is a powerful artificial neural tool to analyse multidimensional data. The standard self-organizing maps have been applied in the economic analysis of multidimensional financial data (Deboeck, 1999). We already used MULTI-SOM for the analysis of multidimensional patent data (Polanco et al, 2001; Lamirel et al, 2001). Here, we apply it to analyse the crucial question "Enterprise and Internet". We say "crucial" because it relates to hot contemporary situation which is today described and analysed within the framework of terms such as "new economy", "knowledge-based economy", "information society" or "network society". All these expressions seek to determine social as well as economic

changes in which one sees as a determining factor the new communication and information technologies.

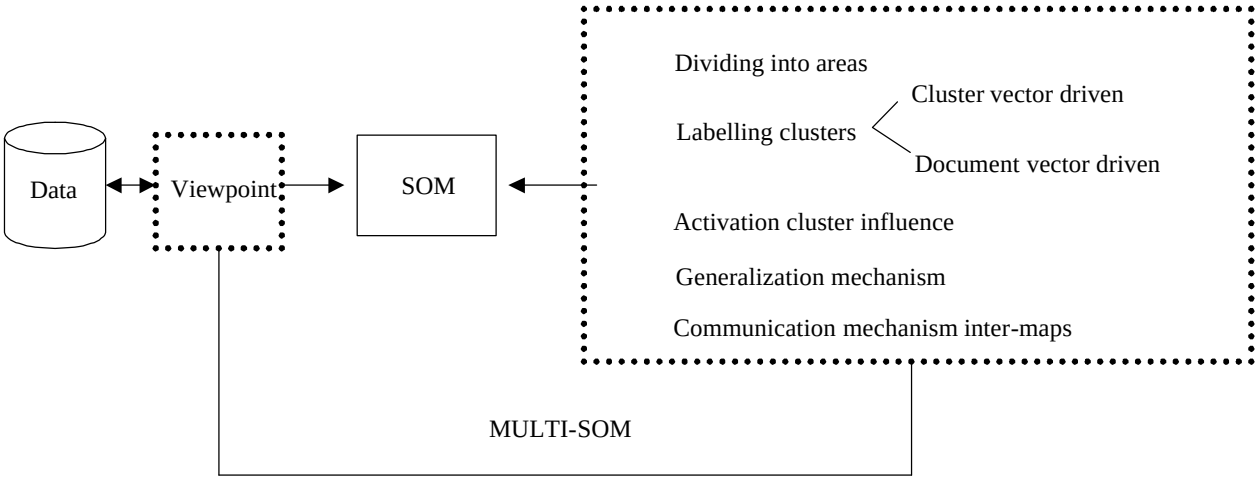


Figure 1 MULTI-SOM components.

The whole represent a model of information analysis. The mechanisms operating on the clusters and the maps are designed to help the task of analysis of information.

As we see on figure 1, the viewpoint is interposed between the data and SOM. Moreover as we have said the maps are organized in areas. In addition, MULTI-SOM offers to the user the following interactive functions at the level of the clusters and maps and their areas. The capacity to act on the labelling of the clusters according to two ways: one is called cluster vector driven and the other document vector driven. The other function of analysis acting on the clusters is to be able to visualize their size and the influence of a cluster on the others. MULTI-SOM also provides to the user the interactive mechanisms on the maps: they are the generalization and communication mechanisms. Generalization is to be regarded as a mechanism of induction making it possible for the user to realize and observe a generalized visualization from particular instances. The communication operates from a map making it possible to see the relations between the elements selected in this map towards the other maps.

The article is organized as follows. In section 2, we present the model implemented in terms of Multi-Self-Organizing Maps. In section 3, we show a real application on a very current subject: “Enterprise and Internet”. In section 4, the conclusions recall the principal components of MULTI-SOM, and the type of analysis that it enables to realize as a contribution to the economic analysis in the information age. The maps representing five points of view on the subject are grouped in the appendix.

2. A MODEL OF INFORMATION ANALYSIS

In this section, the main formal elements will be present. We begin by recalling briefly the SOM clustering process and then the area division principle. Since this model is based on standard self organizing maps shaped in areas or regions. We follow to expose the viewpoint definition, the interactive functions of labelling clusters, and the generalization and communication mechanisms. In our opinion, these last elements represent the most original contributions in the use of SOM.

2.1 Clustering process

SOM realises two basic clustering procedures: selecting a winning node and updating the weights of the winning node and its neighbour nodes.

1. Winning node selection: Let $x(t) = \{x_1(t), x_2(t), \ldots, x_n(t)\}$ be the input profile vector selected at time *t*, and $w_k(t) = \{w_{k1}(t), w_{k2}(t), \ldots, w_{kn}(t)\}$ the reference vector of node *k* at time *t*. The smallest of the Euclidean distances $\|x(t) - w_k(t)\|$ is used to define the winning node *s*: $\|x(t) - w_s(t)\| = \min \|x(t) - w_k(t)\|$.
2. Neighbourhood definition: After the winning node has been selected, the reference vector of *s* and the reference 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 this node again. This is achieved through the following computation: $w_{ki}(t+1) = w_{ki}(t) + a(t) \times h(t) \times [x_i(t) - w_{ki}(t)]$, *for* $1 \le i \le n$; where $a(t)$ is a gain term verifying $0 \le a(t) \le 1$ that decreases in time and converges to *0*, and $h(t)$ is the neighbourhood function.

Once the SOM algorithm is achieved, the data (documents) are affected to the nodes of the map. For each input data, the winning node is selected according to the first step of the algorithm presented above, and the data is affected to this selected node. Nodes who do not beneficiate of any document affectation are called transition nodes. Even if these nodes cannot be considered as document recipients, they play a significant role during learning to maintain the map topographic coherency.

2.2 Dividing map into areas

The cooperative feature of the neighbouring nodes on the map can be used to highlight different closed descriptive areas that have been generated by competitive learning. The size of the resulting descriptive areas on the map displays the relative importance of their associated descriptors in the data set. The neighbouring areas will highlight relationships between their associated descriptors in the data set. The area division algorithm is a dividing and gathering procedure having the capability of managing the areas overlaps. The area computation has been exposed (in Lamirel, 1995) as a generalization of (Lin et al, 1991). The area computation is based on the topographic properties of the node reference vector in SOM. Let the reference vector w_i of each cluster *i* be defined by the set of descriptor weights $\{w_{i1}, w_{i2}, \ldots, w_{in}\}$. The first phase consists in defining the set Q_p such as:

$$Q_p = \left\{ j \mid Ind\left(\underset{i}{Max}\left(w_{ji} \right) \right) = p \right\}$$

where p is a descriptor index, $\underset{i}{Max}\left(w_{ji} \right)$ is the maximal value among all the descriptor weights associated to the profile vector j, and the function $Ind\ \left(w_{ki} \right)$ gives the index of the descriptor associated to the value w_{ki}. Q_p is the set of clusters whose maximal descriptor weight in their profile vectors is related to the descriptor of index *p*. Let *A* be to global set of primitive areas, such as

$$A = \bigcup_{p=1}^{N} \{ Q_p \}$$

The gathering phase is the construction of the set A' of the final areas of the map. The principle of this phase is to gather two areas onto one when other hides the maximal components of an area. It can be described as

$A' = A$

While $\left(\exists Q_p \text{ and } Q_q \text{ in } A' \text{ such as } \forall x \in Q_q, \exists y \in Q_p, n_{yp} \geq n_{xp}\right)$

Do

1) $Q_{p \cup q} = Q_p \cup Q_q$
2) $A' = \left(\left(A - Q_p - Q_q\right) \cup Q_{p \cup q}\right)$

End Do.

Thus, the clusters with the similar dominant descriptors in their reference vector are grouped in the same logical area. The role of the map logical division into spatial areas or clusters groups is to highlight in an overall way the general topics described by the map along with their relative importance.

2.3 Viewpoint definition

In practice, the viewpoint consists in separating the description space of the data set into different subspaces corresponding to different descriptor subsets. Formally, the set *V* of all possible viewpoints issued from the description space *D* of a data set is defined as:

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

v_i represents a viewpoint and $P(D)$ represents the set of the parts of the description space of the data belonging to *D*; the union of the different viewpoints reconstitutes the description space of the data. The viewpoint subsets issued from *V* may be overlapping ones. They also fit into the structure of the data when they correspond to different index subsets associated to the different data sub-fields. Each viewpoint is achieved in the form of a self-organizing map (SOM) shaped in areas or regions. The viewpoint is an absolutely exogenous concept to SOM. On the contrary, the following three elements represent extensions closely related to the structure of SOM

2.4 Labelling clusters

The nodes that are graphically represented by circles on the map display are clusters. The cluster labelling strategies define cluster labels that could optimally represent the clusters contents when the map is displayed. Taking into account the fact that there is no absolute strategy for achieving that goal, the choice has been to implement two different ways that could be used during the map interactive session.

The first labelling way is compliant with the principle of area construction described above (in section 2.2). The label l_i of the cluster *i* is the descriptor *k* associated to the vector component that verifies: $k = \underset{j}{Max}\left(w_{ij}\right)$. The second labelling way is data oriented. It is based on the analysis of the cluster member vectors according to the best member labelling strategy or the average member labelling strategy.

If M_i is the member set of the cluster *i* and *p* the member verifying:

$$\left\| m_p - w_i \right\| = \underset{k \in M_i}{Min} \left\| m_k - w_i \right\|$$

Then the label l_i of the cluster i is the descriptor k associated to the vector component of m_p that verifies: $k = \underset{j}{Max}\left(m_{pj}\right)$

If $\overline{a}_i$ is the vector verifying: $\overline{a}_i = \frac{1}{\left|M_i\right|} \sum_{k \in M_i} m_k$

Then, the label l_i of the cluster i is the descriptor k associated to the vector component of $\overline{a}_i$ that verifies: $k = \underset{j}{Max}\left(\overline{a}_{ij}\right)$

The cluster labelling strategies are useful in providing the user with complementary information for the map cluster content interpretation. Because some important information on a cluster can be better represented in the member vector than in its related cluster vector. This phenomenon is due to the fact that the cluster vectors are drawn from the clustering process of the document indexes while the member vector represent straightforward information from the original data (i.e., clustered and mapping documents).

2.5 Generalization mechanism

The generalization is the mechanism that starting from a map generates new clustering levels of synthesis by progressively reducing the number of clusters. Since the original basic map has been build on the basis of a 2D square neighbourhood between clusters, the transition from one level to another is achieved by choosing a new cluster set in which each new cluster will represent the average composition of a square of four direct neighbours on the original level. Formally, generalization is defined as follows. Let $n \times m$ (n, $m \geq 2$) be the dimensions of the map associated to a given level, the generalization process will produce a next more general level in the form of a $(n - 1) \times (m - 1)$ map. For each new level cluster n, the vector computation formula applies:

$$W_n^{M+1} = \frac{1}{4} \sum_{n_k \in N_n^M} W_{n_k}$$

where N_n^M is the square neighbourhood set on the map M associated to the cluster n of the new map M+1. This can be considered as the determination of all the square centres of a source level for building a new level. This procedure has the advantage of preserving the original neighbourhood structure on the new generated levels. Moreover it ensures the conservation of topographic properties of the profile vectors of the map clusters, and consequently the conservation of the closeness of the spatial areas of clusters in the generalized maps. After the computation of a new generalized map, the original data (documents) are re-affected to their nodes. For each input data, the winning node is selected according to the first step of the Kohonen algorithm (see above section 2.1). As the basic map and its generalization share the same dataset their communication is based on the principles described in the following section.

2.6 Communication mechanism

The communication process between maps operates in three successive steps. At the beginning, the activity is set up directly by the user or by a query formulation on one or several clusters of one or several source map, according to different scalable modalities. In the next step, the activity is transmitted to the data nodes associated to the activated clusters of the

source map (also called down activation). Finally, the activity is transmitted through the data nodes to other maps to which these data are associated (called up activation). This mechanism operates according to possibility or probability modes of computation which the user can choose.

Let A_i^T the activity of the cluster *i* in the target map *T*. Then, A_i^T can be derived from the activity of a source map *S* according to the following formula: $A_i^T = f_{n \in i}(A_{j_n}^S)$
where *n* represents a node associated to a data, j_n its associated cluster on the source map, *f* is a function implementing the relation, which operates according to possibility or probability modes of computation which the user can choose. In the possibility computation of the relation, each cluster inherited of the activity transmitted by its most activated data (or members). The *f* function can be given as:
$f = \underset{n \in i}{Max}(A_n^+) + \underset{n \in i}{Max}(A_n^-)$
where A^+ represents a positive activity value (positive choice), and A^- a negative activity value (negative choice). Positive as well as negative activity can be managed in the same communication process. In the probability computation of the relation, each cluster inherited of the average activity transmitted by its data members, either they are activated or not. The *f* function associated to the probability mode can then be given as:
$f = \frac{1}{\|M_i\|} \sum_{n \in i} A_n$ where $\|M_i\|$ represents the number of data associated to the class *i*.

Moreover, a bias function, *g*, which can be optionally used, modulates the activity transmission from a cluster to a data (down activation), and afterwards from data to a cluster (top activation), considering the pertinence degree of a data to a cluster as attenuation factor for that transmission.

The possibility computation helps the user to detect weak relationships between components (weak signals) existing between clusters belonging to different viewpoints. For the possibility theory (see Dubois & Prades, 1988). The probability computation gives a more reliable measure of the strength of the relation between components, and may then be used to differentiate between strong and weak matching.

In conclusion, the extension of the standard SOM as MULTI-SOM carries out in itself a model of the analysis of information. What we call the model of analysis is above all the organization of the data in viewpoints and their translation into maps cut out in areas, and the operations which were presented in this section. Now we pass to the application. In the following section, we will remain on a level rather general but able to show the advantage to use the viewpoints and the representation of information contained in a set of data in maps organized in areas.

3. APPLICATION ON "ENTERPRISE AND INTERNET" DATA

The objective of this section is to illustrate a certain level of use of MULTI-SOM, and this on a field of obvious interest today for the economic knowledge. One calls "new economy" a set of evolutions and mechanisms such as the emerging and diffusion of new communication and information technologies, in particular Internet, the new goods and services related to these technologies, the incorporation of these new technologies in the production processes of goods and services, including those of the "old economy" (automobiles, chemistry, transport,

...), and the reorganization of the enterprises around more flexible forms (Artus, 2002). The use of the term "knowledge-based economy" deals with a long tendency increasing resources devoted to the production and the transmission of knowledge (education, training, R&D, …), and a major technological event, the advent of new information and communication technologies (Foray, 2000). In the terminology of the European Union institutions, the expression "information society" is largely used. A "network society" has been defined as "a society where the key social structures and activities are organized around electronically processed information networks. And it is about social networks which process and manage information and are using micro-electronic based technologies" (Castells, 2000, 2001).

In this section, we initially will specify the data relating to this question, then the construction of the viewpoints, and finally the knowledge represented by the maps. The interest of this representation is that it is not the mental representation of an individual in particular. It is the "objective" representation, that which arises from the analysed data, allowing us to have a knowledge representation according to the viewpoints and their interrelationships.

3.1 Data

A set of 832 bibliographic records was extracted from the PASCAL database following the equation "Enterprise and Internet." If we consider the classification code indexing the data, we note that 54% of data belong to information science and 38 % to computer science. In fact, Economics is not covered by PASCAL database. The period covered by the data set is 1992-2003. 82% of the data correspond to the years 1997-2002. Year 2003 only represents 0.4%. What are the topics treated by this information? As we are not specialists in the field, we will limit ourselves to notice the possibilities of analysis.

3.2 Viewpoints construction

The viewpoints represent various angles of analysis of the same data set. Their practical construction was as follows. We used an in-home tool for the exploratory analysis of the 832 data - STANALYST (NEURODOC) (Polanco et al, 2001). That enabled us to define the points of view. Once the points of view are defined conceptually remains their manual construction. This one consists in a selection of keywords by point of view. The points of view are then built using the standard SOM package. Thus, they become self-organizing maps. We define the following five viewpoints about the Internet effect within enterprises:

1. Enterprise strategy and management
2. e-commerce or e-business
3. Information (information and computer technologies)
4. Knowledge and learning
5. Technology watching and economic intelligence

The first point of view enables us to visualise and analyse the Internet effects on the strategy and management of the firms. The second concentrates around the concepts of e-commerce and e-business and e-trade The third proposes more particularly to visualize and analyse the role of information within the enterprises. Here, information obeys a double meaning, a computer science and technology meaning, and the other is related to management of contents, namely, as documents. The fourth point of view is the Internet use inside of the enterprises in relation with knowledge, learning and training. The fifth point of view refers to the activities on economic or business intelligence and technology survey.

3.3 The knowledge displayed by the maps

What does one visualize on the maps? From the point of view of their use, the maps can be regarded as visual-based means of analysis. According to this proposal the maps represent tools for discovering and analysing research topics. The maps allow users to evaluate the relative position of topics in the two-dimensional space of representation. Always from the point of view of their use, each neuron or node being in oneself a cluster represents a topic, a theme, a category of data, and their proximities (neighbourhoods) and distances on the map are as many signs of the thematic organization of the knowledge space. The shaping of the maps in logical zones contributes even more to the realization of this analytical use, namely the thematic analysis of the research topics. In the appendix the maps show the areas grouping a number of clusters. The areas allow to visualize the common research topic (or theme) of a subset of clusters on the map.

Map 1 displays the "enterprise management and strategy" viewpoint: What effects the Internet has on the management and strategy of the enterprises? Its use in the production processes and also in the organization structure? The Internet effects on the production processes of the "old" economy, and on the new sector of the communication and information? It is considered that the new information technologies have (and in particular the Internet) strongly modified the nature, organization and behaviour of the enterprises. Map 1 shows 23 thematic areas on this subject.

Map 2 displays the "e-commerce or e-business" viewpoint: The use of the Internet by the companies is only one aspect of the new economy. Another aspect is its development in business, trade and commerce. Map 2 shows 19 thematic areas about this particular Internet effect on the economy.

Map 3 displays the "information and communication" viewpoint: This map-based viewpoint focuses on the information and communication technology (ICT) as a major technological force underlying changes in the enterprises (map 1), commerce or trade and business (map 2). Map 3 shows 25 thematic areas about information and computer topics that are involved in new economy.

Map 4 displays the "knowledge and learning" viewpoint: The new economy is said to be also a "knowledge-based economy". Certainly, knowledge always was in the core of the economic development. But now with the concept of knowledge-based economy, the economists suggest the idea of the increasing importance of knowledge and learning or training in the modes of organization and production having as support the new information and communication technologies (NICT) (see Steinmuller, 2002). The NICTs should contribute to accelerate the knowledge investments. The NICTs radically change the conditions of reproduction and transmission of knowledge. Map 4 shows 25 thematic areas about the Internet effect on learning or training and knowledge within the enterprises.

Map 5 displays the "watching and intelligence" viewpoint: Here what one seeks to know is the contribution or the effect of the Internet on the activities of economic intelligence and technological survey. These activities are based on new information technologies and they are related to competitiveness through the information analysis interesting companies and markets: the transformation of information into useful knowledge. These activities constitute

still another aspect of the new economy (maps 1, 2, 3) as a knowledge-based economy (map 4). In this connection, map 5 shows 24 thematic areas.

4. Summary and Conclusions

Let us recall the main components of MULTI-SOM. The concept of point of view, or viewpoint, and the implementation of the various points of view into self-organising maps shaped into logical zones. The inter-active labelling of clusters, and the zone of influence of a cluster. The generalization mechanism allowing to carry out a topographic induction. And finally, the mechanism of inter-map communication making it possible to follow the inter-map relations at the level of the clusters or thematic areas (i.e., logical zones). All these mechanisms operating on the clusters and maps are designed to help the task of information analysis.

We are able, using MULTI-SOM, to analyse information (i.e., 832 data on "enterprise and internet") according to five points of view, and thus to produce knowledge and to display the thematic organization embedded in the data. We can widen this representation with a sixth point of view which would be that of the countries of the authors affiliations. We limited ourselves to show for each point of view the research topics on the maps. The analysis which it is possible to realize using the interactive mechanisms on the clusters (labelling and zones of influence), and on the maps (generalization and intercommunication) exceed the limits of this article. Nevertheless, we propose the model of analysis that MULTI-SOM allows as a SOM-based contribution interesting the economic analysis in the information age.

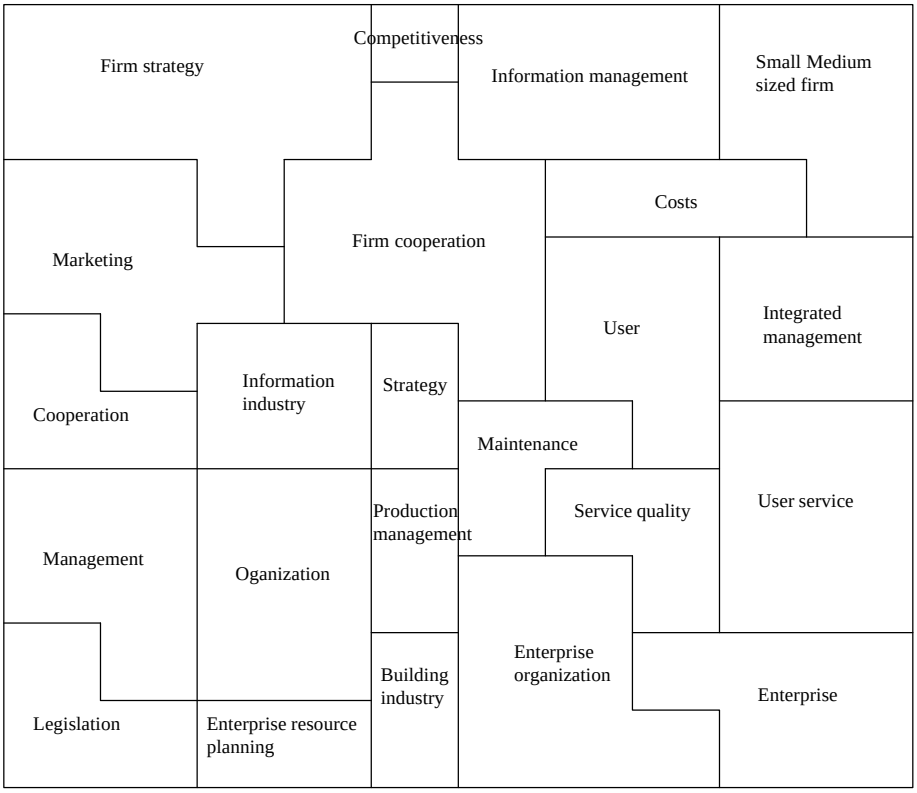
The application of the MULTI-SOM prototype on "Enterprise and Internet" date set can be downloaded at http://eicstes.inist.fr/visualisation_en.html. In this application all the interactive functionalities of the prototype, to which we referred above, and we exposed in section 2 (see also figure 1), can be used.

References

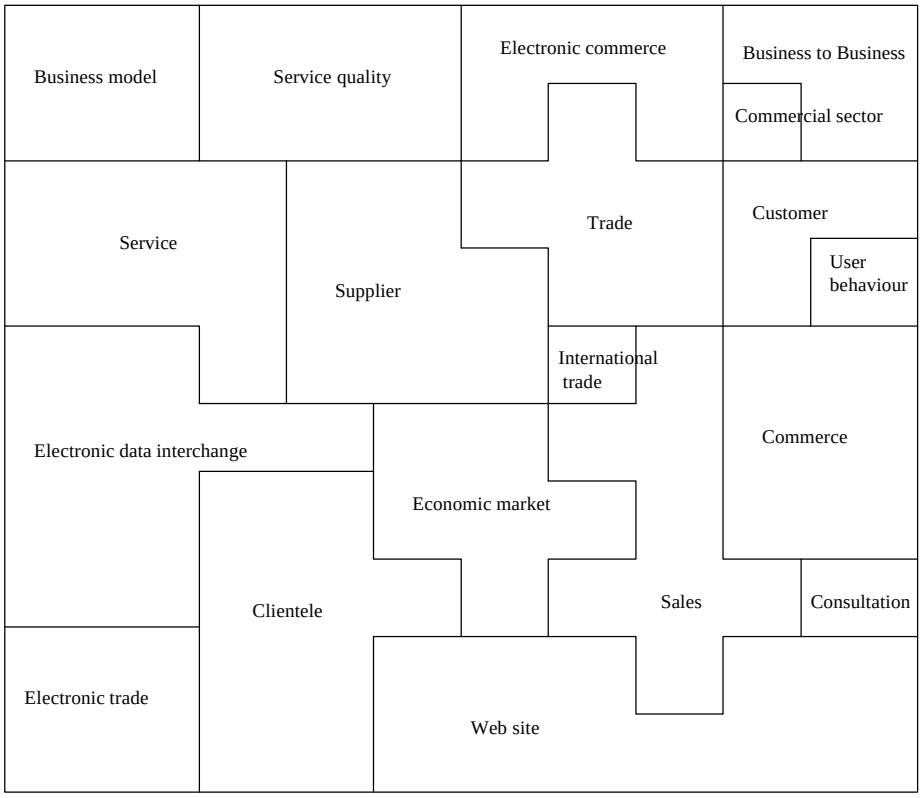
1. Artus P. (2002), *La nouvelle économie*. Paris, La Découverte (collection Repères, vol. 303).
2. Castells M. (2000), *The Rise of the Network Society*. Blackwell
3. Castells M. (2001), http://globetrotter.berkeley.edu/people/Castells/castells-con0.html
4. Deboeck G. J. 1999 Value Maps : Finding value in markets that are expensive, in Oja & Kaski (1999), p. 15-31.
5. Dubois D. and Prades H. (1988), *Possibility Theory*. New York, Plenum Press.
6. Foray D. (2000), *L'économie de la connaissance*. Paris, La Découverte (collection Repères, vol. 302).
7. Kohonen T. (1997), *The Self-Organizing Maps*. Berlin, Springer.
8. Lamirel J-Ch. (1995), *Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif*. Thèse de l'Université de Nancy 1 Henri Poincaré
9. Lamirel J-Ch, Toussaint Y., François C., Polanco X. (2001), Using Artificial Neural Networks for mapping of science and technology : application to patents analysis, *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, July 16-20th 2001, Sydney, Australia. Vol. 1, p. 339-353.

10. Lin X., Soergel D., Marchionini G. (1991) A Self-Organizing Semantic Map for Information Retrieval, *Proceedings of the 4th International SIGIR Conference on R&D in Information Retrieval*, 13-16 October, Chicago, p. 262-269.
11. Oja E and Kaski S. (1999) *Kohonen Maps*. Amsterdam, ELSEVIR.
12. Polanco X., François C., Lamirel J-Ch. (2001) Using artificial neural networks for mapping science and technology : a multi-self-organizing maps approach, *Scientometrics*, vol. 51 (1), p. 267-292.
13. Polanco X., François C., Royauté J., Besagni D., Roche I. (2001) STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology, *Proceedings of the 8th International Conference on Scientometrics and Informetrics*, Sydney, Australia, July 16-20th 2001. Vol. 2, p. 871-873.
14. Steinmueller W. E. (2002) Les économies fondées sur le savoir : leurs liens avec les technologies de l'information et de la communication, *Revue Internationale des Sciences Sociales*, N° 171, p. 159-173

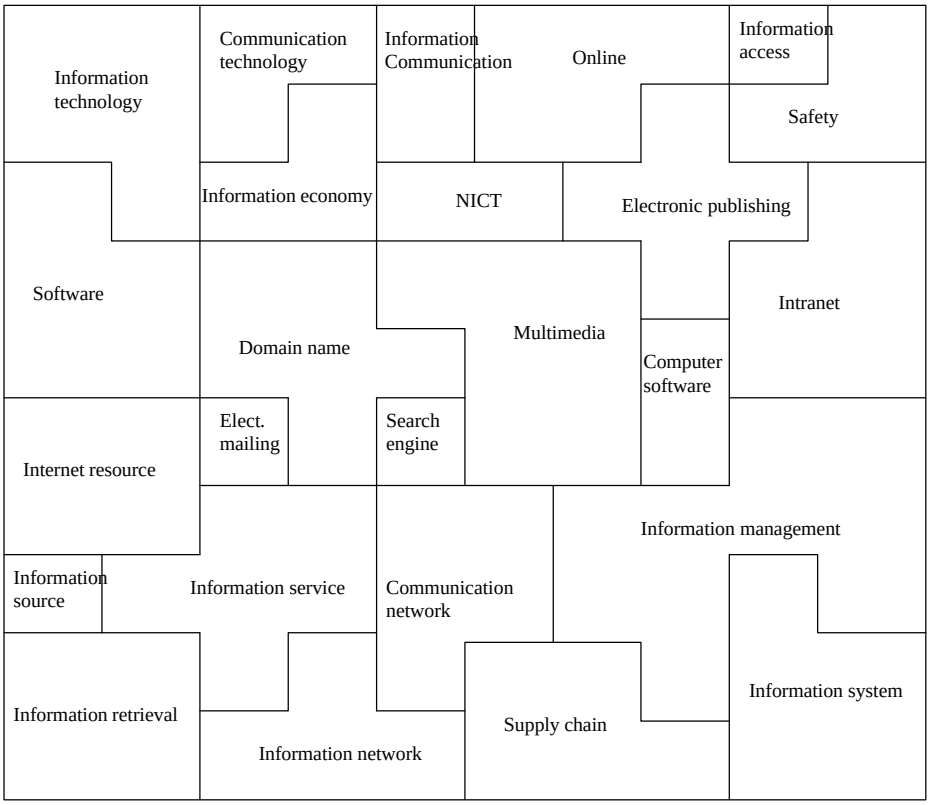
APPENDIX



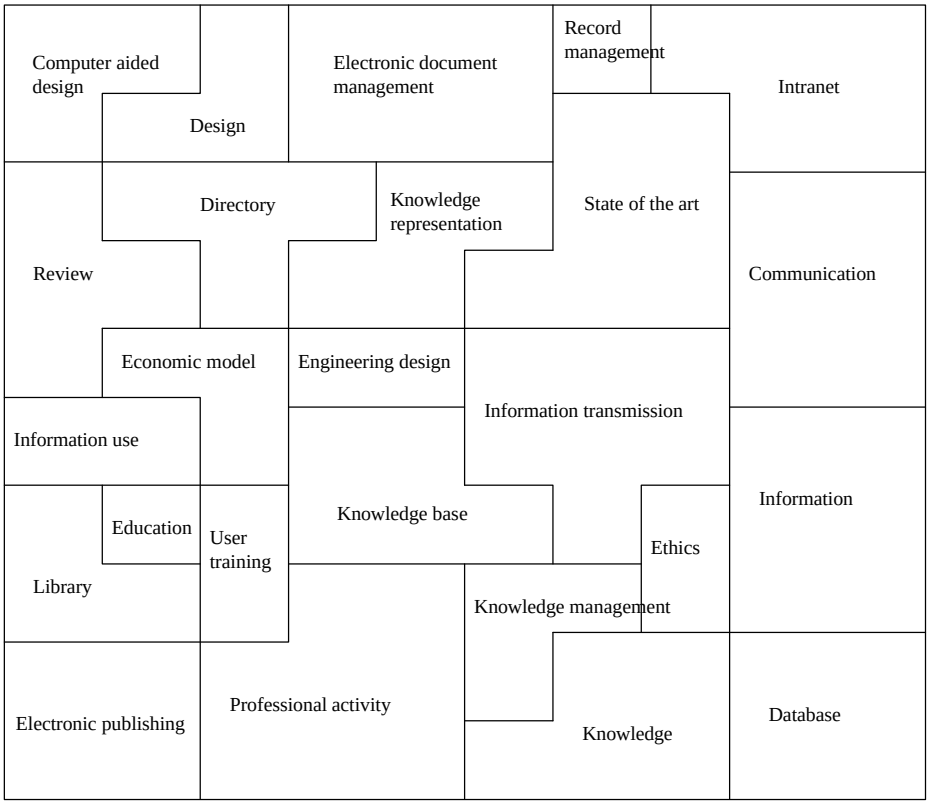
Map 1 Enterprise strategy and management



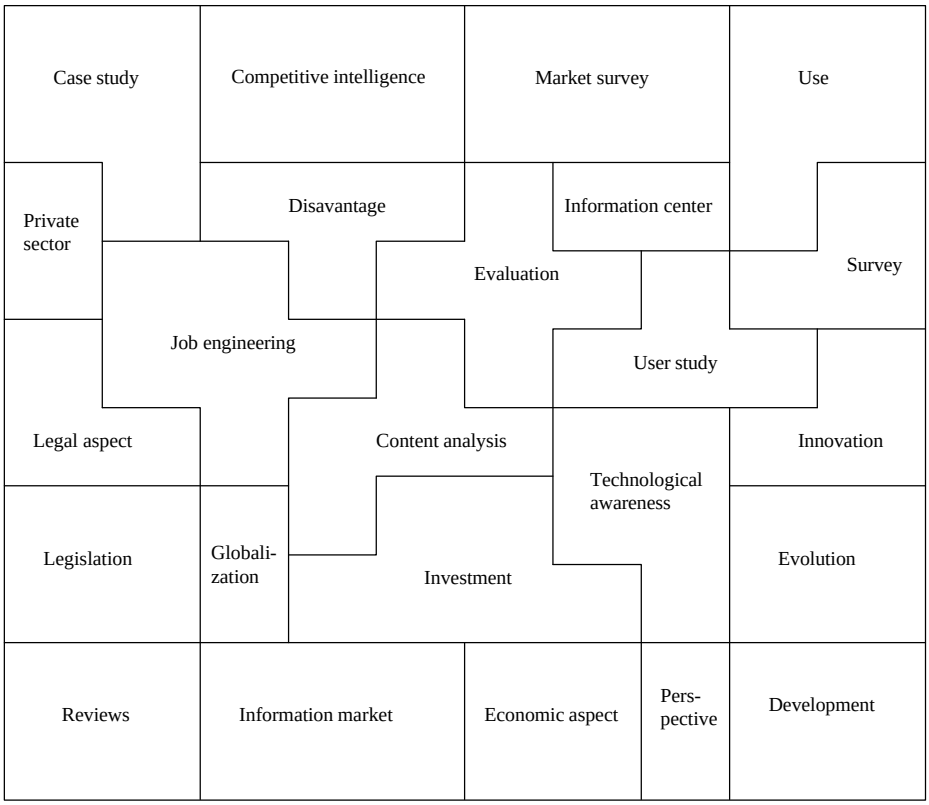
Map 2 E-commerce or e-business



Map 3 Information (information and computer sciences)



Map 4 Knowledge and learning



Map 5 Technology watching and economic intelligence