Evaluation of Collaboration between European Universities using Dynamic Interaction between Multiple Sources

Jean-Charles Lamirel*, Shadi Al Shehabi* and Claire Francois

*lamirel@loria.fr, alshehab@loria.fr
LORIA (Laboratoire lOrrain pour la Recherche en Informatique et ses Applications), Cortex research team, Campus Scientifique, BP 239, F-54506 Vandoeuvre-Lès-Nancy (France)

***Claire.Francois@inist.fr*
INIST (Institut National pour l'Information Scientifique et Technique), URI dept., Allée du parc de Brabois, F-54500 Vandoeuvre-Lès-Nancy (France)

Abstract

This paper presents a new approach whose aim is to extent the scope of Webometrics by providing it with new knowledge extraction and visualization capabilities. An application of this approach on a dataset of websites issued from European universities is presented. The basic model which is considered in this paper is a multi-topographic neural network model. The powerful features of this model are its generalization mechanism and its mechanism of communication between topographies. These two mechanisms allow rule extraction to be performed whenever a single viewpoint or multiple viewpoints on the same data are considered. The association rule extraction is itself based on original quality measures which evaluate to what extent a numerical classification model behaves as a natural symbolic classifier such as a Galois lattice. The visualization of the results of the analyses is based on an original hyperbolic approach

Introduction

Data mining or knowledge discovery in database (KDD) refers to the non-trivial process of discovering interesting, implicit, and previously unknown knowledge from large databases (Han, M. Kamber & Tung, 2001). Such a task implies to be able to perform analyses on high-dimensional input data. The most popular models used in KDD are the symbolic models. Unfortunately, these models suffer of very serious limitations. Rule generation is a highly time-consuming process that generates a huge number of rules, including a large ratio of redundant rules. Hence, this prohibits any kind of rule computation and selection as soon as data are numerous and they are represented by very high-dimensional description space. This latter situation is very often encountered with documentary data.

To cope with these problems, preliminary KDD trials using numerical models have been made. An algorithm for knowledge extraction from self-organizing network is proposed in (Hammer, Rechtien & Strickert, 2002). This approach is based on a supervised generalized relevance learning vector quantization (GRLVQ) which is used for extracting decision trees. The different paths of the generated trees are then used for denoting rules. Nevertheless, the main defect of this method is to necessitate training data. On our own side, we have proposed in (Lamirel, Toussaint & Al Shehabi, 2003) a hybrid classification method for mapping an explicative structure issued from a symbolic classification into an unsupervised numerical self-organizing map (SOM). SOM map and Galois lattice are generated on the same data. The cosine projection is then used for associating lattice concepts to the SOM classes. Concepts properties act as explanation for the SOM classes. Although it establishes interesting bridges between numerical and symbolic worlds this approach necessitates the time-consuming computation of a Galois lattice.

In a parallel way, in order to enhance both the quality and the granularity of the data analysis and to reduce the noise which is inevitably generated in an overall classification approach, we have introduced in Lamirel (1995) the multi-viewpoint analysis based on a significant extension of the SOM model, named MultiSOM. The viewpoint building principle consists in separating the description of the data into several sub-descriptions corresponding different property subsets. In MultiSOM each viewpoint is represented by a single SOM map. The conservation of an overall view of the analysis is achieved through the use of a communication mechanism between the maps, which is itself based on Bayesian inference. The advantage of the multi-viewpoint analysis provided by MultiSOM as compared to the global analysis provided by the classical SOM analysis prososed by Kohonen (2001) has been clearly demonstrated in (Lamirel, Al Shehabi, Hoffman & Francois, 2003)

Proceedings of ISSI 2005, pages 740–749.
Edited by P. Ingwersen and B. Larsen.

for precise mining tasks like patent analysis or Webometrics (Lamirel, Al Shehabi, Francois & Polanco, 2004). Another important mechanism provided by the MultiSOM model is its on-line generalization mechanism that can be used to tune the level of precision of the analysis. Furthermore, we have proposed in (Lamirel & Al Shehabi, 2005) to use the neural gas (NG) model as a basis for extending the MultiSOM model to a MultiGAS model. Hence, NG model proposed by (Martinetz & Schulten, 1991) is known as more efficient and homogeneous than SOM model for classification tasks where explicit visualization of the data analysis results is not required.

In this paper we propose a new approach that consists in using the MultiGAS model both for knowledge extraction and for data analysis visualization purpose in the context of a Webometrics application. The dataset that is used in our experiments is the reference dataset of European websites that has been build up in the framework of the EISCTES project (1999). The knowledge extraction phase consists in using our MultiGAS model as a front-end for unsupervised extraction of association rules. In our approach we exploit both the generalization and the intercommunication mechanism of the model. We also make use of our original recall and precision measures that derive from the Galois lattice theory and from Information Retrieval (IR) domains (Lamirel, Al Shehabi, Francois & Hoffman, 2004). The problem of the visualization of the gas results is solved by the use of an original hyperbolic visualization algorithm. The first section presents the MultiGAS model. The second section presents both the rule extraction principles and the hyperbolic visualisation principles based on the MultiGAS model. The experiment is presented in the last section.

MultiGAS Model

Inter-gas Communication Mechanism

The inter-gas communication mechanism enables to highlight (in an automatic or in an interactive way) semantic relationships between different topics belonging to different viewpoints related to the same data. In MultiGAS, this communication is based on the use of the data that have been projected onto the gas as intermediary neurons or activity transmitters between gases. The intercommunication process between gases operates in three successive steps. The inter-gas communication is established by standard Bayesian inference network propagation algorithm which is used to compute the posterior probabilities of target gas's neurons T_k which inherited of the activity (evidence Q) transmitted by its associated data neurons D. These computations can be carried out efficiently because of the specific Bayesian inference network topology that can be associated to the MultiGAS model. Hence, it is possible to compute the probability $P(act|T_k,Q)$ for an activity of modality act_m on the gas neuron T_k which is inherited from activities generated on the source gas. This computation is achieved as follows (Lamirel, Al Shehabi, Francois & Polanco, 2004):

$$P(act_m|T_k,Q)=\frac{\sum_{D_j\in act_m,T_k} Sim(D_j,S_i)}{\sum_{D_j\in T_k} Sim(D_j,S_i)} \qquad \textbf{(1)}$$

such that the similarity $Sim(D_j,S_i)$ is the cosine correlation measure between the codebook vector of the data D_j and that of the neurone S_i. The neurons of the target gas getting the highest probabilities can be considered as the ones who include the topics sharing the strongest relationship with the topics belonging to the activated neurons of the source gas.

Generalization Mechanism

The main roles of the generalization mechanism are both to evaluate the coherency of the topics that have been computed on an original gas and to summarize the contents of this later into more generic topics. Our NG generalization mechanism presented in (Al Shehabi & Lamirel, 2005) creates its specific link structure in which each neuron of a given level is linked to its 2-nearest neighbours (Fig. 1). For each new level neuron n the following codebook vector computation applies:

$$W_n^{M+1}=\frac{1}{3}\left(W_n^M+\sum_{n_k\in V_n^M} W_{n_k}\right) \qquad \textbf{(2)}$$

where V_n^{M+1} represents the 2-nearest neighbour neurons of the neuron *n* on the level *M* associated to the neuron *n* of the new generated level *M+1.* After codebook vector computation the repeated neurons of the new level (i.e. the neurons of the new level that share the same codebook vector) are summarized into a single neuron. The proposed generalization mechanism can be also considered as an implicit and distributed form of a hierarchical classification method based on neighbourhood reciprocity. Its main advantage is to produce homogeneous generalization levels while ensuring the conservation of the topographic properties of the gas codebook vectors on each level. Moreover, the inter-gas communication mechanism presented in the former section can be used on a given viewpoint between a gas and its generalizations as soon as they share the same projected data.

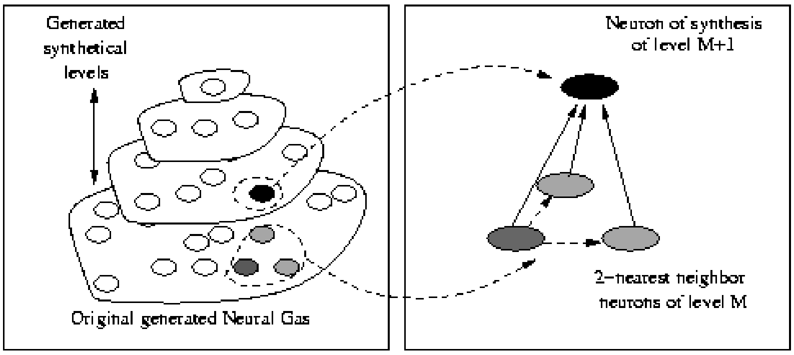
Generated synthetical levels
Original generated Neural Gas
Neuron of synthesis of level M+1
2-nearest neighbor neurons of level M

Figure 1: Gas generalization mechanism (2D representation of gas is used for the sake of clarity of the figure).

Quality of Classification Model

The classical evaluation measures for the quality of classification are based on the intra-class inertia and the inter-class inertia. These measures are often strongly biased because they depend both on the pre-processing and on the classification methods. Therefore, we have proposed in (Lamirel, Al Shehabi, Francois & Hoffman, 2004) to derive from the Galois lattice and Information Retrieval (IR) domains two new quality measures, *Recall* and *Precision*. As compared to classical inertia measures, averaged measures of *Recall* and *Precision* present the main advantages to be independent of the classification method. The *Precision* and *Recall* measures are based on the properties of class members. The *Precision criterion* measures in which proportion the content of the classes generated by a classification method is homogeneous. The greater the *Precision*, the nearer the intensions of the data belonging to the same classes will be one with respect to the other, and consequently, the more homogenous will be the classes. In a complementary way, the *Recall criterion* measures the exhaustiveness of the content of said classes, evaluating to what extent single properties are associated with single classes. The *Recall* (Rec) and *Precision* (Prec) measures for a given property *p* are expressed as:

$$\underset{c}{\mathrm{Re}}\, c(p) = \frac{\left|c_p^*\right|}{\left|C_p^*\right|} \qquad \textbf{(3)}$$

$$\underset{c}{\mathrm{Pr}}\, ec(p) = \frac{\left|c_p^*\right|}{\left|c\right|} \qquad \textbf{(4)}$$

such that, *C* is a set of classes issued from a classification method applied on a set of documents *D*, and

$$c_p^* = \left\{ d \in c, \quad W_c^p > 0 \right\} \qquad \textbf{(8)}$$

We have demonstrated that if both values of *Recall* and *Precision* reach the unity value, the peculiar set of classes represents a Galois lattice. A class belongs to the peculiar set of classes of a given classification if it possesses peculiar properties. Finally, a property is considered as peculiar for a given class if it is maximized by the class members.

Averaged measures of *Recall* and *Precision* can be used for overall comparison of classification methods and for optimisation of the results of a method relatively to a given dataset. In this paper we will more specifically focus on peculiar properties of the classes and on local measures of *Precision* and *Recall* associated to single classes. Hence, as soon as this information can be fruitfully exploited for generating explanations on the contents of individual classes, as it is demonstrated in (Lamirel, Toussaint & Al Shehabi, 2003), it will also represent a sound basis for extracting rules from said classes.

Rules Extraction from MultiGAS Model

An elaborated unsupervised neural model, like MultiGAS, represents a natural candidate to cope with the related problems of rule inflation and rule selection that are inherent to symbolic methods. Hence, its synthesis capabilities that can be used both for reducing the number of rules and for extracting the most significant ones. In the knowledge extraction task, the generalization mechanism can be specifically used for controlling the number of extracted association rules. The intercommunication mechanism will be useful for highlighting association rules figuring out relationships between topics belonging to different viewpoints.

Rules Extraction by the Generalization Mechanism

We will rely on our own class quality criteria for extracting rules from the classes of the original gas and its generalizations. For a given class *c*, the general form of the extraction algorithm (**A1**) follows:

```
$\forall p_1, p_2 \in P_c^*$
1) If (Rec($p_1, p_2$) = Prec($p_1, p_2$) = 1) Then: $p_1 \leftrightarrow p_2$ (equivalence rule)
2) ElseIf (Rec($p_1, p_2$) = Prec($p_2$) = 1) Then: $p_1 \rightarrow p_2$
3) ElseIf (Rec($p_1, p_2$) = 1) Then
      If (Extent($p_1$) $\subset$ Extent($p_2$)) Then: $p_1 \rightarrow p_2$
      If (Extent($p_2$) $\subset$ Extent($p_1$)) Then: $p_2 \rightarrow p_1$
      If (Extent($p_1$) $\equiv$ Extent($p_2$)) Then: $p_1 \leftrightarrow p_2$
$\forall p_1 \in P_C^*$, $\forall p_2 \in P_c - P_c^*$
4) If (Rec($p_1$) = 1) If (Extent($p_1$) $\subset$ Extent($p_2$)) Then: $p_1 \rightarrow p_2$ (*)
```

where Prec and Rec respectively represent the local *Precision* and *Recall* measures, Extent(*p*) represents the extension of the property *p* (i.e. the list of data to which the property *p* is associated), and P_c^* represent the set of peculiar properties of the class *c*.

The optional step 4) (*) can be used for extracting extended rules. For extended rules, the constraint of peculiarity is not applied to the most general property. Hence, the extension of this latter property can include data being outside of the scope of the current class *c*.

Rules Extraction by the Inter-gas Communication Mechanism

A complementary extraction strategy consists in making use of the extraction algorithm in combination with the principle of communication between viewpoints for extracting rules. The general form of the extraction algorithm (**A2**) between two viewpoints v_1 and v_2 will be:

```
$\forall p_1 \in P_C^*$, $\forall p_2 \in P_{C`}^*$ and $C \in v_1$, $C` \in v_2$
1) If (Rec($p_1$) = Rec( $p_2$) = Prec($p_1$) = Prec($p_2$) = 1) Then
      If (Extent$_{v1}$($p_1$) $\subset$ Extent$_{v2}$($p_2$)) Then: $p_1 \rightarrow p_2$ (association rule)
      If (Extent$_{v2}$($p_2$) $\subset$ Extent$_{v1}$($p_1$)) Then: $p_2 \rightarrow p_1$ (association rule)
      If (Extent$_{v1}$($p_1$) $\equiv$ Extent$_{v2}$($p_2$)) Then: $p_1 \leftrightarrow p_2$ (equivalence rule)
2) ElseIf (Rec($p_1$) = Rec($p_2$) = Prec($p_2$) = 1) Then: $p_1 \rightarrow p_2$
3) ElseIf (Rec($p_1$) = Rec($p_2$) = 1) Then
      If (Extent$_{v1}$($p_1$) $\subset$ Extent$_{v2}$($p_2$)) Then: $p_1 \rightarrow p_2$
      If (Extent$_{v2}$($p_2$) $\subset$ Extent$_{v1}$($p_1$)) Then: $p_2 \rightarrow p_1$
      If (Extent$_{v1}$($p_1$) $\equiv$ Extent$_{v2}$($p_2$)) Then: $p_1 \leftrightarrow p_2$
```

Extended rules will be obtained as follows:
$\forall p_1 \in \mathrm{P_C}^*$, $\forall p_2 \in \mathrm{P_{C'}}$
Substituting respectively Rec(p_2) and Prec(p_2) by the *viewpoint-based measures* $\mathrm{Rec}_{v1}(p_2)$ and $\mathrm{Prec}_{v1}(p_2)$, related to the source viewpoint, in the previous algorithm.

Hyperbolic Visualization

NG approaches are known for producing more accurate results than the classical data analysis approaches. Nevertheless, overall visualization of gas results represents a complicated problem. When linear projection methods are used for that purpose, the result is a significant loss of information. When non linear projection methods are used, the loss of information is reduced but the neighbourhood structure between classes cannot be properly visualized. An example of such visualization problems is given on the figure 2. On its own side, hyperbolic visualization is known for its capability to cope with the problem of cognitive overload produced by the graph-based approaches. Hence, it permits to visualize complex relationships between data thanks to the use of a focus and context mechanism. Up to now, hyperbolic visualization has mostly been used for solving data organisation problems, like folder or hyperlink management. The original approach that is presented in this paper consists in combining gas analysis with hyperbolic visualization. Hence, the hyperbolic visualization algorithm that will be used in our experiments is an algorithm that has specially developed for that purpose. Its main advantage is to produce a hierarchical classification of the gas in which the information on data density issued from the original neuron (i.e. class) space is preserved.

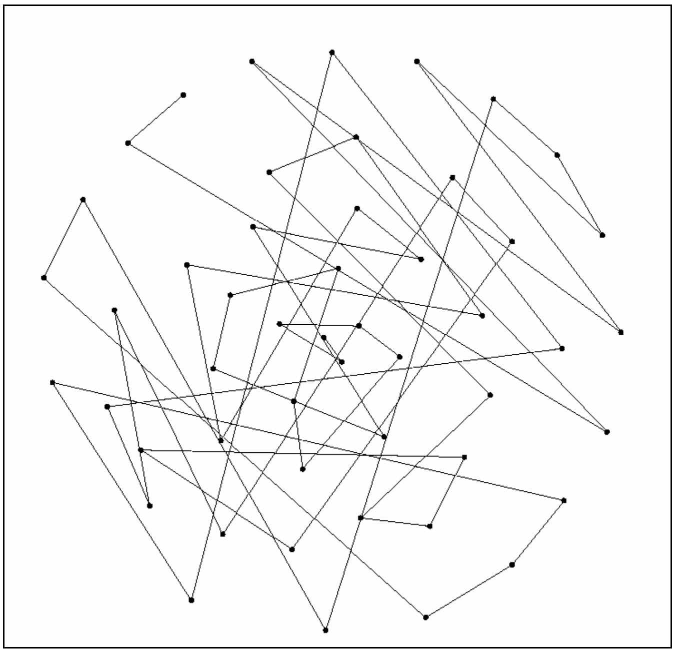

Figure 2: Visualization of neighborhood relations in a 5O neurons (i.e. classes) gas. The complexity of obtained structure is too high and do not permit any interpretation. This problem is typical of cognitive overhead often encountered in graph visualization.

Experimental Results

The data used in our experiment are information about European University websites collected in work-package WP3 of the EICSTES project (1999). The visualized and investigated data set is generated automatically by a tool developed by Isidro F. Aguillo of CINDOC-CSIC Madrid, Spain during EICSTES project and covers the Web sites of universities and research institutes in the 15 countries of the European Union before March 2004. These data had originally the form of multiple tables issued from a preliminary basic structural analysis of the websites. Thus, a first phase of our data preparation consists in obtaining an overall description of each Website, by merging all the individual description tables in one global description table. After that first merge the global description of a Website takes the form of:

- URL
- Name of the organization and department
- Number of pages
- Number of links
- Number of repeated links
- Number of recursive extern links
- Geographic situation : code and town name
- Domain: code, label and related domain codes.

The domain categorization of the websites is based on the UNESCO classification. The UNESCO code is a classification, which is used to allocate a scientific domain to of a Web site starting from its content). As the original dataset is too general, we decided in a first step to focus our study on a specific thematic domain. The selection of data subset related to this domain represents the second step of our data preparation. It is also based on UNESCO code. The 1203 UNESCO code that deals in a global way with computer science and its sub-domains is used for websites extraction. A first context set of 2839 websites is selected this way. In a second step, we choose to focus on German university websites as reference websites for our study. A second kernel set of 378 websites is selected this way. Hence, the study will more precisely focus on the relationships existing between German universities (kernel set) relatively to an European context (context set).

Viewpoints Definition

The different information associated with the selected websites enables us to define 6 different viewpoints: (1) UNESCO codes (**Ucodes**), (2) German universities and their related cities (**Cities**) (3) Links coming from European universities to German universities (**In-Links**), (4) Links starting from German universities to European universities (**Out-Links**), (5) Hyperlinks coupling between German universities based on In-Links (**In-Coupl**), (6) Hyperlinks coupling between German universities based on Out-Links (**Out-Coupl**).

Each viewpoint is represented in the form of a matrix including websites codes in its rows and specific viewpoint criteria in its columns. For each viewpoint complementary tables and indexes are also constructed. The matrices and the indexes will represent the input files for the basic clustering NG application. The tables will be used after the clustering process in the MultiGAS application to provide global information about the generated classes.

Cluster Construction Using Neural Gas Algorithm and Optimisation Criteria

For each viewpoint an optimal gas is calculated. This gas is generated thanks to an optimization algorithm based on the quality criteria we have proposed in (Lamirel, Al Shehabi, Francois & Hoffmann, 2004). Many basic gases of different neuron counts are thus calculated and their qualities are compared. A basic gas is itself represented by the class matrix or node matrix computed by NG basic unsupervised classification algorithm (Martinetz & Schulten, 1991). The basic gas construction is divided in different sub-steps:

a) The data transformation: an IDF conversion according to (Robertson & Spark-Jones, 1976) is applied on the row vectors of the data matrix in order to obtain numeric vectors where the importance of the different descriptors is weighted, followed by a normalization of these numeric vectors.

b) The initialisation of the node matrix: each line of this matrix corresponds to a node on the map. The parameters used are the number of nodes.

c) The computation of the classes: the result of this step which consists in using NG algorithm is a node matrix where each vector representing a node description is updated according to its proximity to the data affected to it and the learning process of its neighbours on the map.

d) The affectation of the websites to the classes: the data (i.e. websites) are affected to the nodes according to their proximity to the nodes description.

The NG output is made of numerical information about the gas: the matrix of the nodes and the document affectations. Each node of the matrix represents a cluster of data. Thus, in a second step, the numerical information about these nodes is associated to information about data affected to classes in a common SGML file. For each node or cluster, this file includes:

- The geographical information used to build the maps: position of the node on the grid defining the map,
- Profile of the node representing a cluster,
- The list of the members with their site identifiers and their URLs.

The files obtained at the end of this basic cluster construction phase are: the global members base and for each viewpoint, a descriptor index file and a cluster file. The summary of this process is given o table 1.

Table 1: Summary of the gas generation process (the highest possible value for the quality factor is 1, **Out-Links** and **Out-Coupl** viewpoints are not managed in our following experiments)

	UCodes	**Cities**	**In-Links**	**Out-Links**	**In-Coupl**	**Out-Coupl**
Dimension	93	96	2839	2839	378	378
Number of classes of gas	50	50	50	-	15	-
Gas quality factor	0,82	0,97	0,64	-	0,41	-

Hyperbolic Visualization of Data Analysis Results

Our first experiment consists in building hyperbolic trees for visualizing the results of the gas analyses of two different viewpoints, that is the **Cities** viewpoint and the **UCodes** viewpoint. The goal of this is to highlight groups of interaction between cities (universities) and topics, respectively. The resulting trees are presented at figure 3 and figure 4.

The two examples of generated tree show that hyperbolic visualization represents a useful tool for interpretation of the results of a data analysis. It main advantage is to suppress the cognitive overload of classical interaction analysis while conserving the most important information. One of our perspectives in a very near future is to apply it to link analysis. Hence, the current version of our algorithm is not optimised for very high dimensional data. That the reason why it has not been yet applied to the results of the In-Links viewpoint.

Knowledge Extraction Thanks to Association Rules

In this first experiment we make use of the intra-viewpoint extraction algo (**A1**), in extended mode, for extracting extended rules from the results of the gas analysis of the single **In-Coupl** viewpoint. The role the rule extraction related to such viewpoint is to find expressions of dependence between German universities through the external view of other Europeans universities. Indeed a rule of the type $A \rightarrow B$ means in such a viewpoint that the university A is viewed (from the European point of view) as a satellite of the university B.

Our algorithm produced very significant results. Hence, 1909 extended rules have been extracted from the gas, including 604 extended rules with a support > 2 (the higher the support and the confidence of a rule, the higher will be its plausibility). The overall results of rule extraction are presented in table 3. Some examples of extracted rules are given hereafter.

www.tu-chemnitz.de → www.sc.rwth-aachen.de (supp = 19, conf = 100%)
www.fh-niederrhein.de → www.tu-chemnitz.de (supp = 5, conf = 100%) : peculiar

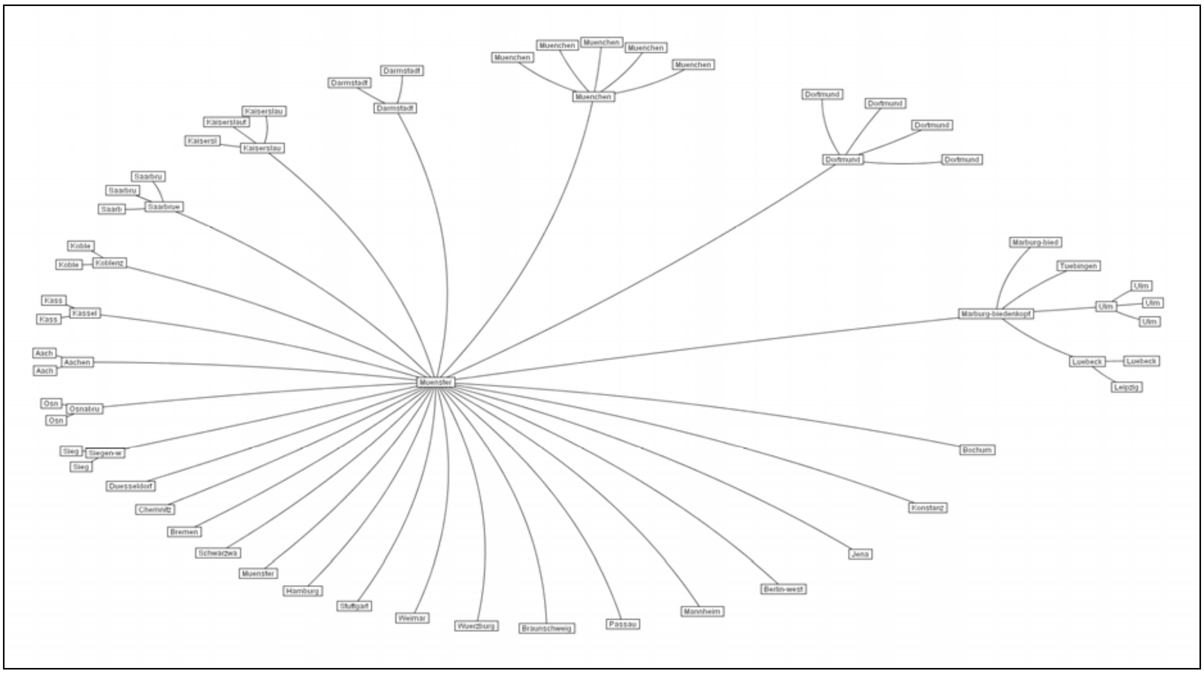

Figure 3 : Hyperbolic tree of the German cities possessing universities. The analysis that can be performed with such a tree is interactive. The focus can be set by the analyst on a specific node of the tree. In this case, the chosen node will shift to the center of the tree and the children of that node will be automatically unfolded. A branch of the tree including many levels corresponds to rich poles of influence. When the label of a node remain stable when one goes up too the root of the tree (initially set at the center of the tree), it corresponds to a very influent topics. The city that has been considered as central for this analysis is the city of **Munster**. That means that research in computer science is especially developed in north Germany.

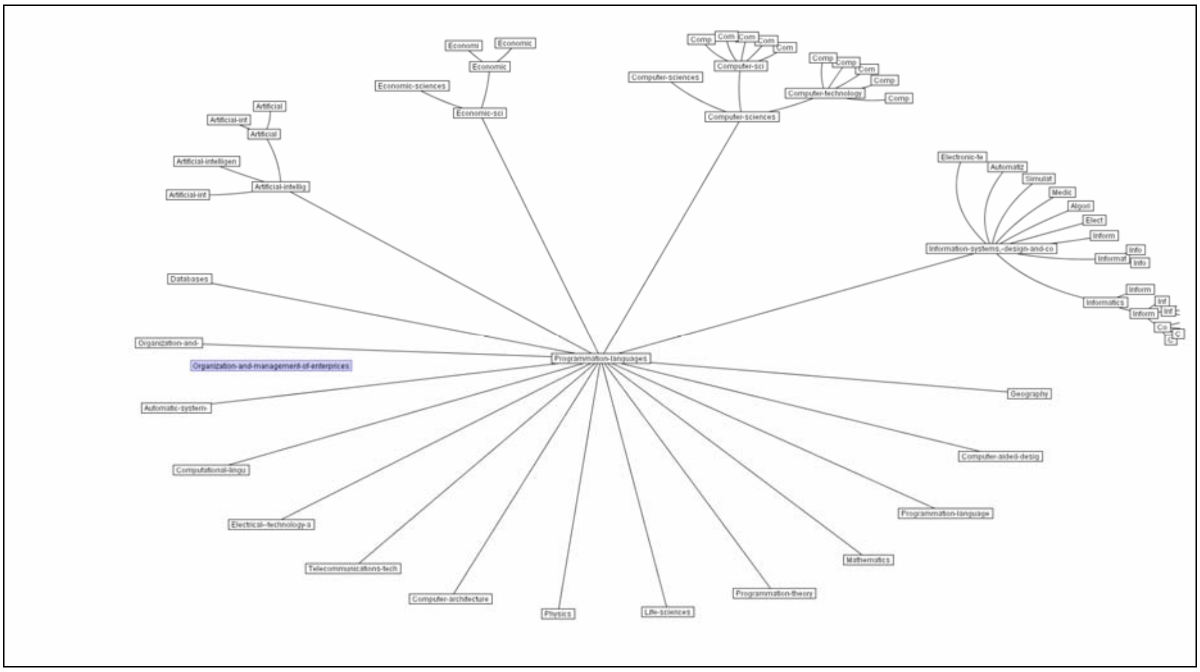

Figure 4: Hyperbolic tree of the research topics managed in German universities. Topics that appear in neighbor branch of the tree are strongly related one to another. As an example, **Economic Sciences** and **Artificial Intelligence** are strongly related in German research, as well as **Databases** and **Organisation of Entreprise**. The **Information systems** topic (at the right of the tree) is also very rich because its include many sub-branches. The central topic related to computer science in Germany is **Programmation languages**. Hence, it represents the root of the hyperbolic tree.

The combination of these two rules highlights a chained dependency between universities viewed from European research context.

In this second experiment we make use of the inter-viewpoints extraction algo (**A2**), in extended mode, for extracting extended rules from the results of the gas analysis of the between the **Cities** and the **UCodes** viewpoints and between **UCodes** and the In-Links viewpoints, respectively. In each case the extraction algorithm is applied is a bidirectional way.

The role the rule extraction related to communication between the **Cities** and the **UCodes** viewpoints is to find expressions of dependence between cities and research topics. Indeed a rule of the type $A \rightarrow B$ *, where A is a UCode and B is a city*, means that the city *A* is specialised in the topic described by *B*. 46 rules (all peculiar) have been extracted from the gas, including 2 rules with a support > 2. The overall results of rule extraction are presented in table 3. An example of extracted rule is given hereafter.

Data transmission. Videofax. Fax. Keys. Cryptography → Saarbruecken (supp = 2, conf = 100%)

The role the rule extraction related to communication between the **In-Links** and the **UCodes** viewpoints is to find expressions of dependence between links and research topics. This rules could be especially used for explaining links. Indeed a rule of the type $A \rightarrow B$ *, where A is a link and B is a Ucode*, means that the link *A* is fully explained by the topic described by *B*. 1002 rules have been extracted from the gas. The overall results of rule extraction are presented in table 3. An example of extracted rule is given hereafter.

www.medizin.uni-koeln.de → Medical science (supp = 3, conf = 100%)

The presented rule give Medical Science as a thematic explanation for all links issued from the *www.medizin.uni-koeln.de* website. This also means that the linking policy of this website is both a homogeneous and closed policy.

Table 1 : Summary of results. The peculiar rule count is the count of rules obtained with the standard versions of the extraction algorithms. The extended rule count is the count of rules obtained with the extended versions of the extraction algorithms including their optional step.

	In-Coupl (intra)	**Cities ↔ UCodes**	**UCodes ↔ Links**
Extended rules (sup > 0)	1909	46	1002
Extended rules (sup > 2)	604	2	Not calculated
Extended rule average confidence	100%	100%	100%
Peculiar rules	4	46	507
Peculiar rule average confidence	100%	100%	100%

The potential of a rule extraction approach for performing accurate webometrics is clearly highlighted by the preceding examples even if a lot of work remains to be done both for sorting the rules and for combining their results in a more synthetic way. In single viewpoint experiment, when our extraction algorithm is used with its optional step, it is able to extract rules count that can be compared to the one obtained with a classical symbolic model which basically uses a combinatory approach. The main advantage of our algorithm, as compared to a classical symbolic method, is the computation time. Indeed, as soon as our algorithm is class-based, the computation time it significantly reduced. Moreover, the lower the generalization level, the more specialized will be the classes, and hence, the lower will be the combinatory effect during computation. Another interesting result is the behaviour of our extraction algorithm when it is used without its optional step. Complementary experiments have shown that in this case, a rule selection process that depends of the generalization level is performed: the higher will be the generalization level, the more rules will be extracted. We have also already done some extension of our algorithm in order to search for partial rules. Our results showed us that, even if this extension is used, no partial rules will be extracted in the low level of generalization when no optional step is used. This tends to prove that the standard version of our algorithm is able to naturally perform rule selection. The results of our multi-viewpoint experiment are similar to the ones of our single viewpoint experiment. A rule selection process is performed when the standard version of our algorithm is used. The maximum extraction performance is obtained when *viewpoint-based Recall* and *viewpoint-based Precision* related to the source viewpoint are used (see algorithm A2).

Conclusion

In this paper we have proposed a new approach for knowledge extraction based on a MultiGAS model. Our approach makes use of original measures of recall and precision for extracting rules from gases. Thanks to the MultiGAS model, our experiments have been conduced on single viewpoint classifications as well as between multiple viewpoints classifications on a reference dataset of European websites that has been build up in the framework of the EISCTES project. In these experiments we have token benefit of the inter-gas communication mechanisms that is embedded in the MultiGAS model. We have also proposed an original algorithm for hyperbolic visualization of the data analysis results. Even if complementary experiments must be done, our first results are very promising. Indeed, they have shown that hyperbolic visualization represents a useful tool for interpretation can results of a data analysis. Moreover, in the webometrics domain, this type of visualization can easily challenge the classical graph-based approach that often produces unmanageable results. The potential of a rule extraction approach for performing accurate webometrics has also been clearly highlighted by our experiments. One of our perspectives is to more deeply develop our model in order to extract rules with larger context like the ones that can be obtained by the use of closed set in symbolic approaches. Another interesting perspective would be to adapt measures issued from information theory, like IDF or entropy, for ranking the rules.

References

Al Shehabi, S. & Lamirel, J-C. (2005). Multi-Topographic Neural Network Communication and Generalization for Multi-Viewpoint Analysis, *Proceedings of International Joint Conference on Neural Networks* (IJCNN'05). Montréal.

Cherfi H. (2004). Étude et réalisation d'un système d'extraction de connaissances à partir de textes. *PhD dissertation*, University of Nancy 1, France.

EISTES project (1999). European Indicator, Cyberspace and the Science-Technology-Economy System. IST-1999-20350.

Hammer B., Rechtien, A. & Strickert, M. (2002). Rule Extraction from Self-Organizing Networks. *Paper submitted to International Conference on Neural Networks* (ICANN 2002).

Han, J., Kamber, M. & Tung, A. K. H. (2001). Spatial clustering methods in data mining: A survey, In H. Miller, H. & Han, J. (Ed.), Geographic Data Mining and Knowledge Discovery: Taylor and Francis.

Kohonen T. (2001). *Self-Organizing Maps*. Berlin: Springer Verlag.

Lamirel, J-C. (1995). Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif. *PhD dissertation*, University of Nancy 1, France.

Lamirel, J-C., Toussaint, Y. & Al Shehabi, S. (2003). A Hybrid Classification Method for Database Contents Analysis. *Proceedings of 16th International Florida Artificial Intelligence Research Society Conference* (FLAIRS 2003). St. Augustine.

Lamirel, J-C., Al Shehabi, S., Hoffmann, M. & Francois, C. (2003). Intelligent patent analysis through the use of a neural network: experiment of multi-viewpoint analysis with the MultiSOM model. *Proceedings of International Conference on Computational Linguistics* (ACL 2003). Sapporo.

Lamirel, J-C., Al Shehabi, S., Hoffmann, M. & Francois, C. (2004). New classification quality estimators for analysis of documentary information: application to web mapping. Scientometrics, 60 (3), 445-462.

Lamirel, J-C., Al Shehabi, S., Francois, C. & Polanco X. (2004). Using a compound approach based on elaborated neural network for Webometrics: an example issued from the EICSTES Project. Scientometrics, 61 (3), 427-441.

Martinetz, T. & Schulten., K. (1991). A "neural-gas" network learns topologies. In Kohonen, T., Mäkisara, K., Simula, O. & Kangas, J. (Ed.), *Artificial neural networks* (pp. 397-402). Amsterdam: North-Holland.

Roberston, S.E. & Spark Jones, K. (1976). Relevance Weighting of Search Terms. Journal of the American Society for Information Science (JASIS), 27, 129-146.