New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping.

Jean-Charles Lamirel*, Claire Francois, Shadi Al Shehabi*, Martial Hoffmann****

*LORIA B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex France
E-mail : lamirel@loria.fr
**URI/INIST-CNRS, 2, allée du Parc de Brabois - 54514 Vandoeuvre-lès-Nancy Cedex France
E-mail : claire.francois@inist.fr

Abstract

The information analysis process includes a cluster analysis or classification step associated to an expert validation of the results. In this paper, we propose new measures for estimating the quality of cluster analysis. These measures derive form the Galois lattice theory and from the Information retrieval (IR) domain. As opposed to classical measures of inertia, they present the main advantages to be both independent of the classification method and of the difference between the intrinsic dimensions of the data and those of the clusters.

We present two experiments using the MultiSOM model, which is an extension of the Kohonen's SOM model, as cluster analysis method. Our first experiment on patents data shows how such measures can be used to compare viewpoint-oriented classification method, like MultiSOM, with a global cluster analysis approach, like Kohonen's SOM. Our second experiment, which takes part in the EISCTES EEC project, highlights that break-even points between our different measures of Recall/Precision can be used in order to determine an optimal number of clusters for Web data classification. The contents of the clusters obtained when using different break-even points are compared in order to study the quality of the resulting maps. This optimisation seems to be mandatory when one want to classify documents issued from the Web, where sparseness is usually a blocking factor.

1. Introduction

In the procedure of information analysis, one general problem is the evaluation of the results of data cluster analysis methods. The complexity of the studied topics combined with the weaknesses of the most widespread objective classification quality estimators, like inertia, may finally led to make use of an expert of the studied domain for a subjective evaluation of the quality of the classification results. In this paper we propose new objective quality estimators for both evaluating and optimising the results of the cluster analysis and of the mapping methods, especially when they are applied in the domain of documentary databases. We have experienced our estimators in two different ways. The first way consists in using them for comparing the efficiency of the viewpoint's oriented data analysis methods with the efficiency of the global analysis methods on the same set of data, composed of a patent collection. The second way consists in using it for optimising the classification results which has been obtained from a large non-homogeneous set of web pages.

2. A new set of measures for cluster quality evaluation

When anyone aims at comparing cluster analysis methods, he will be faced with the problem of choice of reliable classification quality measures. The classical evaluation measures for the

Jean-Charles Lamirel, Claire François, Shadi Al Shehabi and Martial Hoffmann, *New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 129-141.

quality of a cluster analysis are based on the intra-cluster inertia and the inter-cluster inertia (Lebart *et al.* 1982; Rahm, 1980).

The intra-cluster inertia can be defined as:

$$\frac{1}{|C|}\sum_{c\in C}\frac{1}{|c|}\sum_{d\in c}\left\|p_c - p_d\right\|^2 \quad \text{(Eq. 1)}$$

where C represent the set of clusters associated to the classification, d represent a cluster member and p_x represent the profile vector associated to the element x.

The inter-cluster inertia can be defined as:

$$\frac{1}{|C|^2 - |C|}\sum_{c\in C}\sum_{c'\in C, c'\neq c}\left\|p_c - p_{c'}\right\|^2$$

On the basis of these two measures, a classification is considered as good if it possesses low intra-cluster inertia as compared to its inter-cluster inertia. However, these measures are often biased in several ways.

One first bias of these measures is related to the fact that the intrinsic dimensions of the cluster's profiles (number of non-zero components in the profiles) are not of the same order of magnitude than the intrinsic dimensions of the data profiles. It is especially true in the documentary domain, where the average number of indexes in the documents is extremely low as compared to the dimension of their overall description space. This phenomenon causes an abnormal increase of the intra-cluster inertia. Normalisation of the cluster profiles during the cluster construction phase, which is provided by spherical cluster analysis models like Axial K-means (Lelu & François, 1992; Lelu 1993), could indirectly help to solve this problem. Nevertheless an undesirable side effect of this solution is the averaging of the cluster profiles.

The intra-cluster inertia is measuring the distance of the cluster elements from the profile of a cluster to which they have been affected after the classification process. As a second bias, this latter measure might not be able to properly distinguish coherent clusters from incoherent ones. In fact, for the same value of inertia, the elements might be spread around the cluster profile in incoherent clusters, as well as they might be grouped together at a given distance of the cluster profile in coherent ones. To partly cope with this problem a measure of intra-cluster inertia, which only depends of the profile of the cluster elements, has been proposed by (Ould Mohamed Yaha 1997). It is equivalent to:

$$\frac{1}{N}\sum_{i\in C, i=1}^{n-1}\sum_{j\in C, j=i+1}^{n}\left\|p_{d_i} - p_{d_j}\right\|^2, \quad N = \sum_{c\in C}\frac{n_c(n_c-1)}{2} \quad \text{(Eq. 2)}$$

where n_c represents the number of elements of the cluster c.

As the other inertia criteria, this last measure, even if it is completed by median estimation, can only give an average estimation of the coherency, mainly because it manages the cluster member profiles in a global way.

Moreover, the inertia criteria values also strongly depend on the classification methods, making it difficult to achieve comparative efficiency studies between different classification methods. For example, in Kohonen's SOM cluster analysis methods, the constraints generated by the topography building principle tend to make artificially decreasing the inter-cluster inertia as compared to other cluster analysis methods.

Lastly, the inertia measures are not user-oriented because they do not focus on the easiness of interpretation of the cluster contents which is mainly linked with the homogeneity of the

cluster member descriptions. One user-oriented measure will more specifically quantify to which extent the members of a cluster share a common property set instead of focusing on a global distance between member profiles.

Thus, the alternative evaluation measures we proposed in this paper are derived from the Galois lattice theory (Lamirel & Toussaint, 2000) and from Information Retrieval (IR). In our approach, a Galois lattice, *L(D,P)*, is a conceptual hierarchy built on a set of documents *D* which are described by a set of properties *P*. A class of the hierarchy, also called "formal concept", is defined as a pair *(d,p)* where *d* denotes the extension of the concept, i.e. a subset of *D*, and *p* denotes the intention of the concept, i.e. a subset of *P*.

	d_1	d_2	d_3	d_4	d_5	d_6
WEB	×	×	×	×		
Information Retrieval(IR)	×	×	×		×	×
Search Engine(SE)	×	×	×		×	×
Directory(D)	×			×		
Yahoo!(Y)	×					
AltaVista(AV)					×	
Google(G)					×	

Figure 1: Document indexation table

Considering the document indexation table presented hereafter, the resulting Galois lattice that could be generated on the basis of this table is given at the Figure 2.

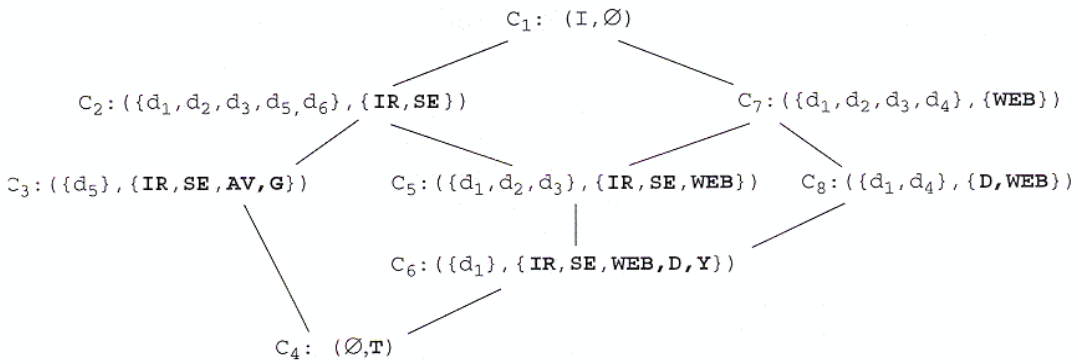


Figure 2: Galois Lattice corresponding to **Fig. 1** table

A Galois lattice could be considered as a pure natural elementary classifier. Indeed, it groups the data into classes by directly considering their intrinsic properties (i.e. without any preliminary construction of class profiles). Hence, we propose to derive from its behaviour news class quality evaluation measures which can validate the intrinsic properties of the numerical clusters. For the sake of user-orientation, our measures will be based in a parallel way on the recall and precision criteria which are extensively used from evaluating the result quality of the information retrieval (IR) systems.

In IR (Salton, 1971), the **Recall *R*** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of relevant documents which should have been found in the documentary database. The **Precision *P*** represents the ratio between the number of relevant documents which have been returned by an IR system for a given query and the total number of documents returned for the said query. **Recall** and **Precision** generally behave in an antagonist way: as **Recall** increases, **Precision** decreases, and conversely. The ***F*** function has thus been proposed in

order to highlight the best compromise between these two values (Van Rijsbergen, 1975). It is given by:

$$F = \frac{2(R * P)}{R + P} \qquad \text{(Eq. 3)}$$

Based on the same principles, the **Recall** and **Precision** measures which we introduce hereafter evaluate the quality of a classification method by measuring the relevance of its resulting cluster content in terms of shared properties. In our further descriptions, the cluster content is supposed to be represented by documents, and the descriptors (i.e. the properties) of the documents are supposed to be weighted by values within the range $[0,1]$.

Let us consider a set of clusters C resulting from a classification method applied on a set of documents D, $\overline{C}$ represents the peculiar set of clusters extracted from the clusters of C, which verifies:

$$\overline{C} = \left\{ c \in C \,\middle|\, S_c \neq \varnothing \right\}.$$

The set of properties S_c which are peculiar to the cluster c is described as:

$$S_c = \left\{ p \in d, d \in c \,\middle|\, \overline{W}_c^p = Max_{c' \in C}\left(\overline{W}_{c'}^p\right) \right\},$$

where :

$$\overline{W}_c^p = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in C} \sum_{d \in c'} W_d^p},$$

and W_x^p represents the weight of the property p for element x.

The **Recall** measure is expressed as:

$$R = \frac{1}{\left|\overline{C}\right|} \sum_{c \in \overline{C}} \frac{1}{S_c} \sum_{p \in S_c} \frac{\left|c_p^*\right|}{\left|C_p^*\right|},$$

where :

$$C_p^* = \left\{ d \in c \,\middle|\, W_d^p \neq 0 \right\}.$$

The **Precision** measure is expressed as:

$$P = \frac{1}{\left|\overline{C}\right|} \sum_{c \in \overline{C}} \frac{1}{S_c} \sum_{p \in S_c} \frac{\left|c_p^*\right|}{|c|}.$$

Similarly to IR, the **F-measure** (described by Eq. 3) could be used to combine **Recall** and **Precision** results.

The **Precision** measures in which proportion the content of the clusters generated by a classification method is homogeneous. The greater will be the **Precision**, the nearer the intentions of the documents belonging to the same clusters will be one to another, and consequently, the easier will be the interpretation of the cluster contents by a user. In a complementary way, the **Recall** measures the exhaustiveness of the content of said clusters, evaluating to what extent peculiar properties are associated to single clusters.

Recall and Precision may be used for evaluating the capability of the numerical classification to act as an information retrieval system. This capability is particularly important if one consider that classification can be the thematic front-end of a documentary database. Indeed it can act as a filter that evaluates the relevance of the user queries relatively to the database contents. Moreover, we can demonstrate that if both values of **Recall** and **Precision** reach the unity value, the peculiar set of clusters $\overline{C}$ represents a Galois lattice. Therefore, the combination of this two measures enables to evaluate to what extent a cluster analysis model can be assimilated to a Galois lattice natural classifier.

3. Comparison between viewpoint oriented and global classification approaches

Our first experiment consists in comparing the viewpoint-oriented classification approach to a global classification approach, in terms of classification quality. For this purpose we will use the quality criteria that we have previously proposed.

The viewpoint building principle consists in separating the description space of the documents into different subspaces corresponding to different keyword or document subsets. The viewpoint-oriented classification principle consists in generating as many classifications as viewpoints, while conserving an overall view on the interaction between the data through an inter-classification communication mechanism. This principle has been extensively described in a preceding paper, which presented the original ***MultiSOM*** model (Lamirel *et al.* 2001; Lamirel, 1995; Polanco *et al.* 2001), a significant extension of the ***Kohonen's SOM*** model (Kohonen, 1997; SOM papers).

Our test database consist 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 four different subfields corresponding to four different viewpoints: **Patentees, Use, Advantages and Title**. When it is full text, the content of the extracted fields associated with the different viewpoints is parsed by a lexicographic analyser in order to extract viewpoint specific indexes. For each specific viewpoint the resulting descriptor set is weighted by means of an ***IDF*** weighting scheme (Robertson, & Sparck Jones, 1976). and a map of 10x10 neurons (clusters) is finally generated. Two global maps representing global cluster analysis, of the ***WebSOM*** type (Kaski *et al.* 1998), of the patents are also constructed. The descriptor sets of these maps represent the union of the descriptor sets of all the specific viewpoints. They only differ one to another by the number of their clusters. The first one (**GlobMin**) is constrained to have the same number of clusters as the viewpoint maps (i.e. 100 clusters). The second one (**GlobMax**) is constrained to have to sum of the number of clusters of all the viewpoint maps (i.e. it becomes a 20x20 map comprising 400 clusters). Furthermore, an ***Axial K-means*** (***AKM***) cluster analysis (Lelu & François, 1992; Lelu 1993) of the **Use** viewpoint is also generated for the comparison of cluster analysis methods. The evaluation results of the Inertia, Precision and Recall analysis are presented in Tables 1 and 2.

By considering the measures of inertia of SOM classifications, the classification of the Patentees is the only one that has significantly lower intra-cluster inertia as compared to its inter-cluster inertia. As it is partly related to the fact that the dimension of the descriptive space (NFI) of this viewpoint is considerably lower than the ones of the other viewpoints, this criterion can only be used for descriptive space of same order of magnitude. Taking this constraint into account, GlobMax seems to be slightly better than GlobMin, and Use classification slightly better than Advantages classification. The comparison between the AKM cluster analysis and the SOM classification on the Use viewpoint highlights the fact that inter-cluster inertia and intra-cluster inertia based on cluster profiles (Eq. 1) have better values for AKM while intra-cluster inertia based on member profile (Eq. 2) has better value

for SOM. These results might confirm the hypothesis that the constraints generated by the topography building principle provided by the SOM method have a tendency to abnormally affect the cluster profiles while not affecting the cluster content coherency. The overall results of the Table 1 might also lead to conclude to the lack of stability of inertia criteria for a reliable classification method comparison.

	Patentees (*MSOM*)	**Title (*MSOM*)**	**Use (*MSOM*)**	**Use (*AKM*)**	**Advantages (*MSOM*)**
NFI	32	589	234	234	207
NMC	28	55	57	32	61
IntraI	0,12	0,78	0,49	0,31	0,62
IntraI(2)			0,54	0,74	
InterI	0,64	0,14	0,29	1,01	0,23
	GlobMin (*MSOM*)	**GlobMax (*MSOM*)**			
NFI	1075	1075			
MNC	89	258			
IntraI	0,81	0,60			
InterI	0,13	0,26			

Table 1: Summary of the results of the inertia evaluation: NFI = number of final indexes, NMC= number of map clusters with members. Note that the NFI of the "global viewpoint" (i.e. 1075) is less than the sum of the NFIs of all the specific viewpoints (i.e. 1089) because there are similar indexes occurring in different viewpoints. InterI = Inter-cluster inertia, IntraI = Intra-cluster inertia computed with Eq. 1, IntraI(2) = Intra-cluster inertia computed with Eq. 2.

	Patentees (*MSOM*)	**Title (*MSOM*)**	**Use (*MSOM*)**	**Use (*AKM*)**	**Advantages (*MSOM*)**
R	0,94	0,89	0,78	0,97	0,77
P	0,92	0,40	0,63	0,55	0,60
F	0,93	0,55	0,70	0,70	0,67
	GlobMin (*MSOM*)	**GlobMax (*MSOM*)**			
R	0,87	0,84			
P	0,48	0,65			
F	0,61	0,68			

Table 2: Summary of the results of Recall and Precision evaluation: The nearer the different values are from 1, the better are the classification results. The F value provides a synthesis of the results of R and P.

As for the inertia criteria, the classification of the Patentees is the only one that has significantly high F value. This measure changes in the same way for the others viewpoints, with a lower amplitude. It highlights the overall superiority of the viewpoint-oriented approach as compared with a global approach with the same number of cluster (GlobMin). As the number of clusters is strongly increased in the global approach (GlobMax), its F value is simultaneously increased, but the advantage of the viewpoint-oriented approach remains obvious in the average : most of F-values of viewpoints are higher than F-value of GlobMax, with a more reasonable number of clusters per maps from a user point of view. The specific case of the Title classification should be discussed here. The bad quality of this classification

is both due to the index sparseness of this field and to an inappropriate number of clusters, relatively to the size of its associated description space. An unbalance between Recall and Precision (in the favour of Recall) can be observed in the case of the worse classifications (GlobMin and Titles). Such an unbalance means that documents with different properties sets are grouped in the same clusters, leading conjointly to the risk of confusion in the interpretation of the content of the clusters by the user. In the case of AKM-use classification method the unbalance between Recall and Precision could be explained by the principle of overlapping clusters on which this method is based.

Our first experiment enables us both to highlight the relative stability of our quality criteria for classification comparison and to demonstrate the superiority of the viewpoint-oriented classification approach as compared with the global classification approach. The quality analysis clearly shows that the viewpoint-oriented approach enhance the easiness of interpretation of a classification by both reducing the number of cluster to be consulted by the user on each viewpoint and providing him with more coherent and exhaustive clusters in terms of content.

4. Optimisation of classification results

A second experiment consists in optimising a cluster analysis of Web pages issued from an institutional site by the study of the joined evolution of our quality criteria thanks to the number of clusters. The optimisation of the cluster number of a classification is a particularly crucial problem in the case of Web pages analysis, mainly because of the sparseness that is inherent to Web data. This experiment takes part in the EISCTES EEC project, whose global objective is to define tools and methods for cartography of the Web. The MultiSOM model takes part in the set of reference models of the project.

The studied Web site is the one of the Computer Laboratory of Cambridge University. The whole site consists in 1353 HTML pages. The first step consisted in collecting the Web pages and then extracting both textual and outgoing links information from each page. Extracted textual information has been then parsed by a lexicographic analyser in order to build the index (Lamirel *et al.* 2001) of the pages content. After human validation, this index resulted in 1230 different terms. The extracted outgoing links have been standardized by keeping only the root of the URL designating the Web site of the links. For example, the original link “http://raw.cs.berkeley.edu/texpoint/index.htm” is transformed into “http://raw.cs.berkeley.edu”. After standardization, this index resulted in 1912 different Web site roots.

Two different viewpoints are then defined, the first one being based on the link information and the second one being based on the indexed textual information. For each specific viewpoint the resulting descriptor set is weighted up by the use of an IDF weighting scheme and different maps are finally generated from 6x6 to 24*24 neurons (clusters).

The principle of our algorithm of classification optimisation, which is presented in Appendix, is to search for a break-even point (i.e. intersection point) between Recall and Precision. This technique is also used extensively in IR for evaluating the IR system quality (salton, 1971). In our own case, the classification which is the nearest from the break-even point will then represent the best compromise between exhaustiveness and specificity. Our algorithm also includes a refinement phase if a first break-even point has been found. The goal of this phase is either to increase the global quality of the result (better F value) or, it this latter operation failed, to decrease the number of clusters for a better cluster readability. When no break-even

point can be found, our algorithm searches for the classification that minimizes the difference between Recall and Precision.

The Figure 1 and 2 shows the evolution of the different criteria, defined above, according to the growth of the number of clusters.

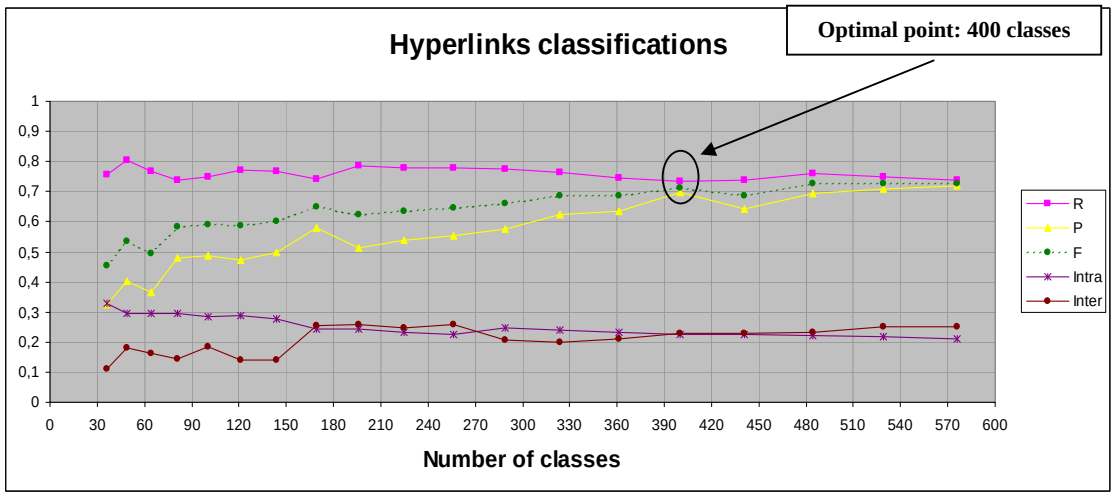


Figure 1: Evolution of the quality measures according to the number of clusters for the hyperlink viewpoint (R = Recall, P = Precision, Intra = Intra-cluster inertia computed with Eq. 1, F = F-value, Inter = Inter-cluster inertia).

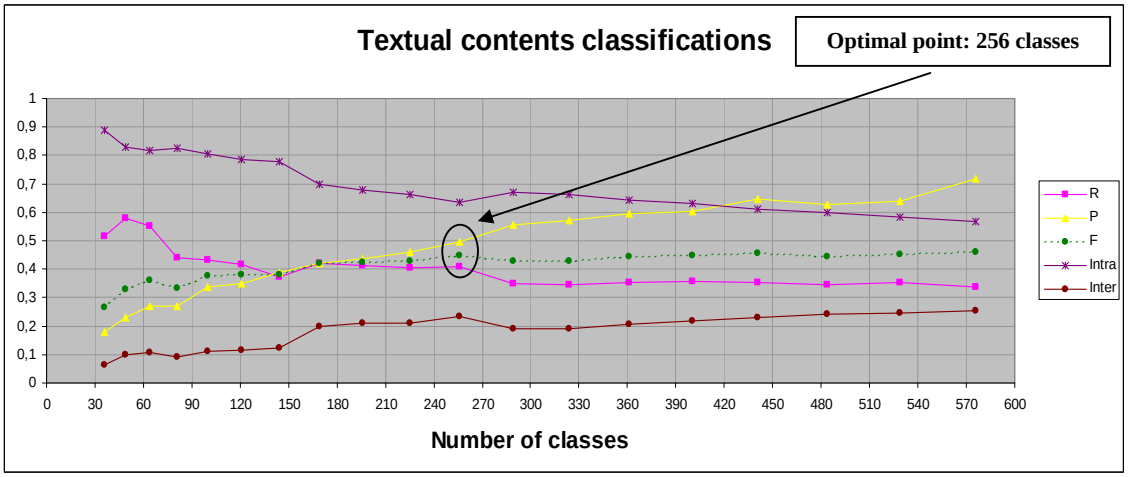


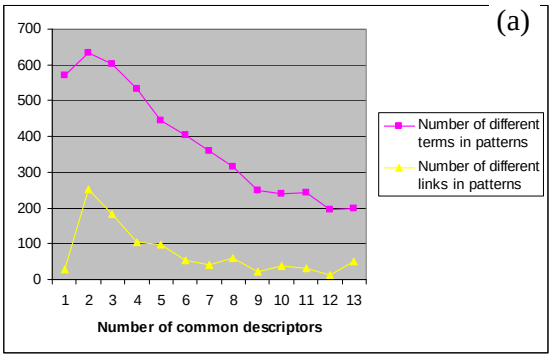
Figure 2: Evolution of the Quality measures according to the number of clusters for the textual content viewpoint (R = Recall, P= Precision, Intra = Intra-cluster inertia computed with Eq. 1, F = F-measure, Inter = Inter-cluster inertia).

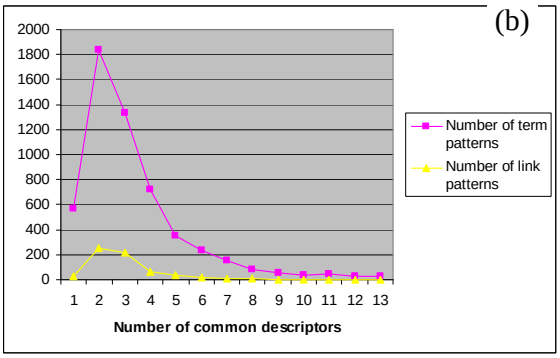
On the Figure 1, the intra and inter-cluster inertia stay in low values, whatever the number of clusters. Conversely, while the Recall measure decreases slowly from 0.8 to 0.7, the Precision measure increases conjointly from 0.3 to 0.7 with an inflexion point at 400 clusters. Our classification optimisation algorithm which is based and the R, P, F measures precisely found this last point as the optimal one. Figure 1 also clearly highlights that this point is a specific point of the R, P, F curves, while characteristic points would be more difficult to automatically highlight on the basis of inertia curves because of numerous intersections between these curves in a constant value range.

On Figure 2, the intra-cluster inertia slowly decreases from 0.9 to 0.6 while the inter-cluster inertia slowly increases from 0.1 to 0.25, and thus no intersection could be found between these curves. The behaviour of inertia curves illustrates the low quality of the classification results. Nevertheless, even in this case, the R, P, F curves present characteristic points. The R and P curves intersect at 144 classes, while an optimal point is found by our algorithm at 256 classes. These two characteristic points are also characteristic points of the inertia curves.

The F curves lookup put into evidence a significant difference of quality between the textual and the link viewpoints, in favour of the link cluster analysis. These differences can be explained through the analysis of the distribution of textual and link indexes in the Web pages. For that purpose, we compute 3 different analyses based on the common descriptors (terms or links) shared by the documents (see Figure 3). In all analyses the x-axis is representing the size of the descriptor patterns by increasing values. The first analysis (a) highlights the number of different descriptors implied in every size of pattern. The second analysis (b) focuses on the number of different combinations of descriptors for each pattern size. The third analysis (c) describes the number of documents which are associated in each pattern size.

The two curves resulting from the first analysis (a) have the same shape for the links and for the terms, with an important peak for patterns of size 2. Nevertheless, the dispersion of the terms appears much more important than the one of the links because of the highest number of terms implied in the patterns. Moreover, the second analysis (b) shows that the richness of pattern combinations is also much more important for the terms than for the links. The third analysis (c) highlights that, when the size of pattern increases, the number documents implicated is much more stable with terms than with links. Quite all the pages (1066 pages among 1353) seems to be linked one to another by a single link (567000 page combinations with only 27 links implicated). This phenomenon could be considered as a bias of our experiment. Indeed, for the sake of simplicity, we choose to only extract the root in the links information. Thus, the linking information between the pages of the main site becomes uniform because the precise destination of the links has been lost.





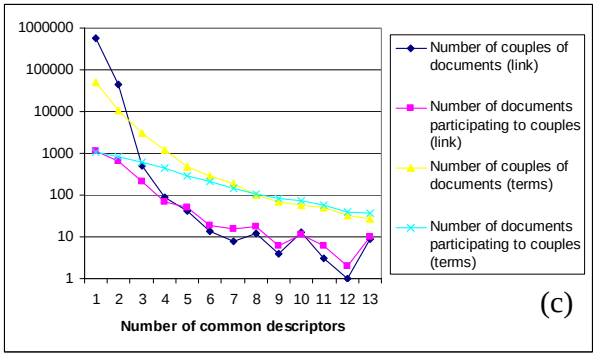


Figure 3: Distribution of terms and links in the Web pages ((a) = Number of descriptors implied in document descriptor patterns, **(b)** = Number of different descriptor pattern implied in documents descriptor pattern, **(c)** = Number of couples of documents/Number of documents being indexed by pattern).

All these criteria confirm that the overlapping in document descriptions is more important when terms are considered. These could explain the fact that discrimination of documents into separate classes is less efficient in textual classification leading to lower cluster quality.

For evaluating the accuracy of our algorithm of classification optimisation, we choose to study the textual viewpoint, as it is easier to analyse. We thus decided to subjectively compare the (16x16=256 clusters) classification which as been considered as the best one by our algorithm with another cluster analysis of similar size (12x12=144 clusters) but of lower quality. As illustrated in Figure 4, the 16x16 map provides the user with more precise information: smaller thematic area on average with more precise labels. As an example, the Figure 4 illustrates in which way two important teams of the Cambridge Computer Laboratory interact. These are the Opera Group working on Distributed Systems and the Rainbow Group working on Human Computer Interaction. The 16x16 map highlights directly the interaction between these groups by their proximity on the map and, moreover, clarifies specific interaction topics through intermediary clusters (i.e. the user interface, interdisciplinary design and computer speech/internet technology classes). In the 12X12 map these two groups are both more distant and separated by two large and general scope clusters.

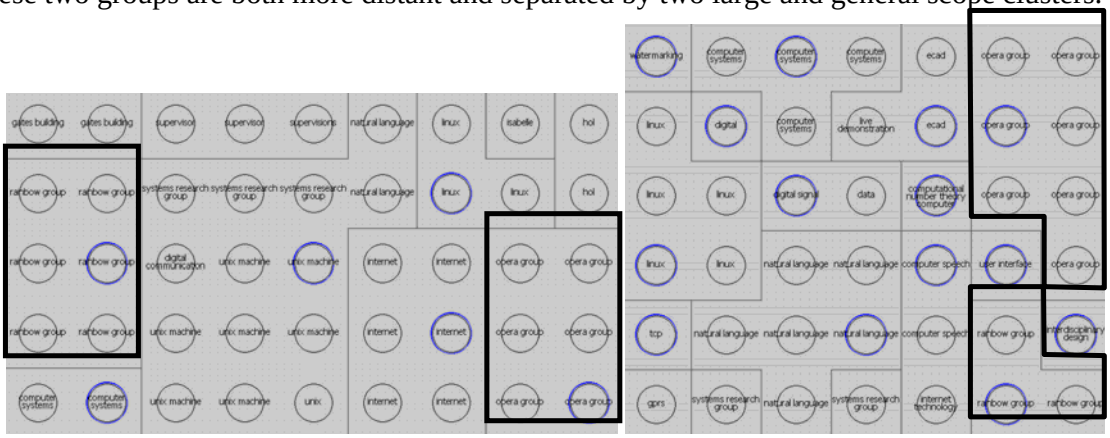


Figure 4: **Comparison between a 12x12 "textual viewpoint" thematic map and a 16x16 "textual viewpoint" thematic map through map extracts:** the 12x12 map extract is presented at the left, the 16x16 map extract is presented at the right. On a map, the names of the clusters illustrate the themes (considering the chosen viewpoint) that have been highlighted by the learning. After the learning, the neurons related to the same themes have been grouped into coherent areas thanks to the topographic properties of the map. The number of neurons of each area can then be considered as a good indicator of the theme weight in the database. The surrounding circles represent the centres of gravity of the areas. For the sake of readability the **Opera Group** and **Rainbow Group** zones are specifically surrounded by thick black lines.

5. Conclusion

In this paper we have proposed new original measures for estimating the quality of cluster analysis. These measures, derived form the Galois lattice theory, present the main advantages to be both independent of the classification method and of the difference between the intrinsic dimensions of the data and those of the clusters, as opposed to classical measures of inertia. They are thus particularly well suited for the evaluation of the quality of documentary classification. Our first experiment showed how such measures can be used to compare viewpoint-oriented classification method, like MultiSOM, with a global cluster analysis

approach, like WebSOM. The experiment results permit us to highlight the superiority of the MultiSOM viewpoint oriented method, which tends to reduce the noise generated by an overall classification process.

Our second experiment highlighted that break-even points between our different measures of Recall/Precision can be used in order to determine an optimal number of clusters for the Web data classification. The contents of the clusters obtained when using different break-even points are compared in order to study the quality of the resulting maps. This optimisation seems to be mandatory when one want to classify documents issued from the Web, where sparseness is usually a blocking factor.

Acknowledgement

The second experiment of this paper is supported by the "European Indicators, Cyberspace and the Science-Technology- Economy System" (EICSTES) project which is a research project (IST-1999-20350) funded by the Fifth Framework Program of R&D of the European Commission. The website of the project is available at: http://www.eicstes.org , the French mirror is available at: http://eicstes.inist.fr/miroir.

Bibliography

Computer Laboratory of Cambridge University Website: http://www.cl.cam.ac.uk

Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). WEBSOM-self organizing maps of document collections. Neurocomputing, 21, 101-117.

Kohonen, T. (1997). 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. Thèse de l'Université de Nancy 1 Henri Poincaré, France.

Lamirel, J.C., Toussaint, Y. (2000). Combining Symbolic and Numeric Techniques for Digital Libraries Contents Classification and Analysis. Proceedings of First DELOS Network of Excellence Workshop, Zurich, December 2000.

Lamirel, J.C., 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 (ISSI), Sydney, Australia, July 2001.

Lebart, L., Morineau, A. & Fénelon, J. P. (1982). Traitement des données statistiques. Paris : Dunod.

Lelu, A. (1993) Modèles Neuronaux pour l'Analyse de Données Documentaires et Textuelles. Thèse de l'Université de Paris 6.

Lelu, A. & François, C. (1992). Information retrieval based on a neural unsupervised extraction of thematic fussy clusters. Neuro-Nîmes 92 : Les réseaux neuro-mimétiques et leurs applications. 2-6 novembre 1992, Nîmes, France

Ould Mohamed Yahya, M.A. (1997). Comparaison de méthodes neuronales avec des méthodes d›analyse des données dans le cadre d'ingénierie de l'information. Mémoire de stage de D.E.S.S. en «Ingénierie mathématique et outils informatiques», Centre Elie Cartan, Université de Nancy 1 Henri Poincaré, France.

Polanco, X., Lamirel, J.C. & François, C. (2001). Using Artificial Neural Networks for Mapping of Science and technology: A Multi self-organizing maps Approach. Scientometrics, 51, 1, 267-292.

Rham (De), C. (1980). La classification hiérarchique ascendante selon la méthode des voisins réciproques. Les cahiers de l'analyse de données, 5, 2, 135-144.

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

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

SOM paper. http://www.cis.hut.fi/nnrc/refs/

Van Rijsbergen, C. J. (1975). Information Retrieval. London : Butterworths.

Appendix: Optimal classification search algorithm

Let S being a sequence of classification indexes,
Suc(S,i) being the function associating to an element *i* its successor in *S*,
Pre(S,i) being the function associating to an element *i* its predecessor in *S*,
S_0 the first element of sequence *S* verifying $Pre(S_0,i)=\varnothing$,
S_* the last element of *S* verifying $Suc(S_*,i)=\varnothing$,
The function R(i),P(i) and *F(i)* computing resp. the Recall, the Precision and the F-value of a classification *i*.

Let S' be the sequence of classification indexes verifying:
$\forall i \in S', NbClass(Suc(i,S')) = NbClass(i) + 2i + 1$

Let S'' be the sequence of classification indexes verifying:
$\forall j \in S'', j \in S' \text{ and } R(j) >= R(Suc(j,S''))$

$Found = False,$
$k = S''_0,$
$While(k \neq \varnothing)\ do$ /* Research of **break-even point** between *R* and *P* */
  $If\ (|R(k) - P(k)| <= 0.05)\ then\ \{Found = True, k_0 = k, k_f = k_0, breakWhile\}$
  $k = Suc(S'', k),$
$EndWhile$
$If\ (Found = True)\ then\ \{$
  $If\ (P(Suc(S', k_0)) > P(k_0))\ then\ \{$ /* Ascending refinement phase (for better global result quality)*/
    $Max_F = F(k_0), R_0 = R(k_0),$
    $k = Suc(S', k_0),$
    $While(k \neq \varnothing)\ do$
      $If\ (F(k) > Max_F)\ then$
        $If\ (R(k) - R_0 <= 0.02)\ then\ \{Max_F = F(k), k_f = k\}\ else\ BreakWhile,$
      $k = Suc(S', k)$
    $EndWhile$
  $\}\ else\ \{$ /* Descending refinement phase (for lower cluster count) */
    $F_0 = F(k_0),\ Min_D = |R(k_0) - P(k_0)|,$
    $k = Pre(S', k_0),$
    $While(k \neq \varnothing)\ do$
      $If\ ((|R(k) - P(k)| <= Min_D)\ and\ (F(k) - F_0 <= 0.02))\ then\ \{Min_D = |R(k) - P(k)|, k_f = k\},$
      $k = Pre(S', k)$
    $EndWhile$
$\}\ else\ \{$ /* Degenerated result: no break-even point found */
  $k = S'_0, Min_D = |R(k_0) - P(k_0)|,$
  $k = Suc(S', k_0),$
  $While(k \neq \varnothing)\ do$
    $If\ (|R(k) - P(k)| <= Min_D)\ then\ \{Min_D = |R(k) - P(k)|,\ k_f = k\},$
    $k = Suc(S', k)$
  $EndWhile$
$\}$
Result $= k_f$

A revised version of this contribution has been published as:

Jean-Charles Lamirel, Claire Francois, Shadi Al Shehabi and Martial Hoffmann (2004). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. Scientometrics, 60(3), 445-462.