Analyse comparative de classifications : apport des règles d'association floues

Pascal Cuxac, Martine Cadot, Claire François

▶ **To cite this version:**

Pascal Cuxac, Martine Cadot, Claire François. Analyse comparative de classifications : apport des règles d'association floues. 5èmes journées d'Extraction et Gestion des connaissances (EGC), CRIP5-SIP – Université René Descartes Paris 5, Jan 2005, Paris, France. pp.519-530. hal-00337092

HAL Id: hal-00337092
https://hal.science/hal-00337092

Submitted on 6 Nov 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse comparative de classifications : apport des règles d'association floues.

Pascal Cuxac*, Martine Cadot**, Claire François*

* INIST-CNRS, 2 allée du Parc de Brabois, 54 154-Vandoeuvre-lès-Nancy Cedex
pascal.cuxac@inist.fr ; claire.francois@inist.fr
** UHP/LORIA, Département d'informatique, BP 239, 54 506- Vandoeuvre-lès-Nancy
martine.cadot@loria.fr

Résumé. Notre travail s'appuie sur l'analyse d'un corpus bibliographique dans le domaine de la géotechnique à l'aide de cartes réalisées avec la plateforme Stanalyst®. Celui-ci intègre un algorithme de classification automatique non hiérarchique (les K-means axiales) donnant des résultats dépendant du nombre de classes demandé. Cette instabilité rend difficile toute comparaison entre classifications, et laisse un doute quant au choix du nombre de classes nécessaire pour représenter correctement un domaine.
Nous comparons les résultats de classifications selon 3 protocoles : (1) analyse des intitulés des classes ; (2) relations entre les classes à partir des membres communs ; (3) règles d'association floues.
Les graphes obtenus présentant des similitudes remarquables, nous privilégions les règles d'association floues : elles sont extraites automatiquement et se basent sur la description des classes et non des membres. Ceci nous permet donc d'analyser des classifications issues de corpus différents.

1 Introduction

Les méthodes de cartographie de l'information sont devenues indispensables pour analyser de gros corpus de données bibliographiques dans le cadre de besoins de veille scientifique ou d'analyses stratégiques de la recherche. Elles s'appuient sur la combinaison de techniques de classification automatique et de cartographie. Pour réaliser ces analyses, nous utilisons la plateforme Stanalyst® (Polanco et al. 2001) qui permet de traiter des corpus bibliographiques et inclut la méthode des K-means Axiales comme méthode de classification.

Cette méthode des K-means axiales (Lelu 1993, Lelu et François 1992, Polanco et François 2000) est basée sur le principe de classification par centres mobiles, plus connue sous le nom de K-means, voir Forgy (1965) pour sa variante non adaptative et MacQueen (1967) pour sa variante adaptative. Comme ces méthodes, la méthode des K-means axiales forme des groupements d'individus en affectant les éléments à des classes provisoires, puis en recentrant ces classes, et en recommençant ces deux phases de façon itérative. Cependant, cette méthode réalise une analyse factorielle sphérique sur chaque classe, les classes sont donc matérialisées par des demi-axes représentatifs des éléments. L'utilisation de cet axe permet de quantifier l'appartenance d'un élément à une classe (typicité). De plus, au lieu d'affecter l'élément à la seule classe où sa valeur est la plus grande, on l'affecte également aux classes pour lesquelles cette valeur dépasse un certain seuil. Il est alors possible

d'obtenir des classes recouvrantes. Cet algorithme, paramétré par le nombre de classes désiré et le seuil des coordonnées des éléments et descripteurs sur les axes, permet donc de construire des classes **recouvrantes** où les individus et descripteurs (documents et mots-clés) sont **ordonnés** selon un degré de ressemblance au type idéal de la classe.

Comme toutes les méthodes d'agrégation autour des centres mobiles, cette méthode converge vers des optima locaux. Une partition optimale peut être approchée soit par l'utilisation d'une stratégie mixte alternant classification de type centres mobiles et classification hiérarchique sur le même jeu de données, soit par la procédure de recherche des groupements stables ou « formes fortes » (Diday 1972). De même, la combinaison de critères de qualité est utilisée pour la détermination d'un nombre optimum de classes (Lamirel et al. 2004). Cette démarche reste délicate car ces critères numériques sont assez pauvres sémantiquement. Plutôt que de chercher une partition optimale, nous proposons donc de conserver plusieurs classifications et de définir les outils nécessaires pour comparer les résultats. La concordance entre deux classifications peut être mesurée en utilisant un indice de type indice de Rand (Youness et Saporta, 2004). De même, l'utilisation de tableaux croisés des éléments des classes permet de décrire les résultats des différentes classifications possibles. Cependant, notre but à long terme étant d'appréhender l'influence du paramètre temps sur l'évolution d'une thématique, il est nécessaire d'avoir une méthode d'analyse dépendant des **caractéristiques des classes (mots-clés)** et non des **éléments (documents)** qui peuvent changer.

Les règles d'association floues définies par Cadot et Napoli (2004) mesurent un lien entre deux objets, en se basant sur leurs caractéristiques. Elles peuvent être appliquées sur la description des classes obtenues par la méthode des K-means axiales : en effet elles mesurent les liens entre classes en s'appuyant sur les valeurs floues (entre 0 et 1) et non binaires (0 ou 1) des caractéristiques des classes.

Le but de ce travail est d'évaluer le potentiel qu'offrent ces règles d'association floues pour la description et la comparaison des différentes classifications. Dans cet article, nous décrivons donc les 3 protocoles de comparaison de classification que nous utilisons : (1) analyse de l'intitulé des classes ; (2) graphes décrivant les relations entre les classes à partir des membres communs ; (3) graphes décrivant les relations entre les classes à partir des règles d'associations floues. Puis, nous comparons et analysons les résultats obtenus, afin d'estimer si la méthode automatique d'extraction des règles d'association floues peut donner des résultats qui remplacent avantageusement ceux obtenus à partir des tableaux croisés des documents.

2 Méthodologie

Le corpus traité, avec la plateforme Stanalyst®, est constitué de 3203 notices bibliographiques extraites de la base PASCAL sur le thème de la géotechnique, publiées en 2001 et 2002 et indexées manuellement. Pour des soucis de clarté des résultats et des analyses nous avons calculé quatre classifications avec la méthode des K-means axiales en fixant successivement le nombre de classes à 20, 30, 40 puis 50. Dans la suite de l'article elles sont nommées respectivement Ndoc20, Ndoc30, Ndoc40 et Ndoc50.

Dans cette partie, nous présentons les trois protocoles de comparaison utilisés. Le premier est uniquement basé sur les intitulés de classes et leur présence ou leur absence dans les classifications. Le deuxième s'appuie sur le nombre de documents communs aux classes de deux classifications successives. Le troisième enfin utilise les règles d'association floues.

2.1 Méthode basée sur les intitulés de classes

La première approche pour comparer nos quatre classifications consiste à étudier les intitulés de classes. Les classes obtenues sont décrites par des mots-clés **ordonnés** selon un indice de typicité dont la méthode de calcul est donnée par Lelu (1993). La pondération utilisée permet de faire ressortir les termes spécifiques de la classe. Une classe est alors nommée automatiquement par le mot-clé de « typicité » la plus forte. En comparant les différents résultats, nous observons que les intitulés de classes peuvent être présents dans toutes les classifications mais aussi apparaître ou disparaître quand le nombre de classe augmente. Cette approche très simple, permet de repérer les classes stables mais manque de précision quant aux modifications des classes entre deux résultats issus de paramétrages différents.

2.2 Méthode basée sur les documents communs

La deuxième approche permet d'affiner l'étude et se base sur l'analyse du nombre de documents communs entre deux classes de deux classifications différentes. Pour cela nous calculons les tableaux croisant les effectifs de deux classifications, avec pour chaque case le nombre de documents communs entre classes. Chaque valeur du tableau est ensuite divisée par la somme des valeurs de la ligne concernée, définissant un indice de force (I.F.). Afin de faciliter une représentation graphique des relations entre classes, l'indice de force est discrétisé selon le protocole ci-dessous :

XXXX si valeur = 1 : on a alors une classe complètement incluse dans l'autre

XXX si 0,75 ≤ valeur < 1 : relation très forte (nombre élevé de documents communs)

XX si 0,5 ≤ valeur < 0,75 : relation forte

X si 0,25 ≤ valeur < 0,5 : relation moyenne

x si 0,1 ≤ valeur < 0,25 : relation faible

A partir de l'ensemble de ces tableaux croisés nous dessinons un graphe faisant apparaître les relations interclasses, où l'épaisseur de la flèche est fonction du symbole défini ci-dessus et où le nombre de relations au départ d'une classe est fonction du nombre de cases renseignées dans la ligne concernée dans le tableau. La figure 1 illustre de façon schématique notre protocole.

Classification M / Classification N	Classe E	Classe F	Classe G	Classe H	Classe I
Classe A	XXXX				
Classe B		XXX		x	
Classe C		x	X	XX	
Classe D			X		X

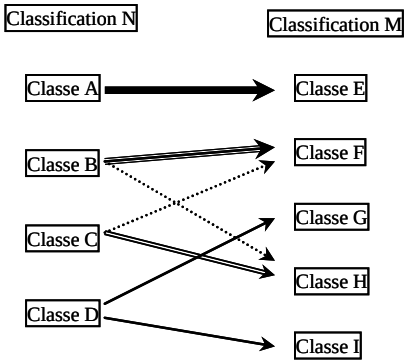

FIG. 1 – *Construction d'un graphe à partir d'un tableau croisé.*

2.3 Méthode basée sur les règles d'association floues

Une règle d'association A→B extraite d'une base de données, représente un lien établi entre deux ensembles de propriétés A et B de cette base de données (Han et Kamber, 2001), lien dont la qualité est évaluée d'après les effectifs des objets de la base les vérifiant. Pour mesurer la qualité de cette règle, on dispose de nombreux indices (Kodratoff, 2001) fonctions de ces effectifs, dont les plus courants sont le support, qui est le nombre d'objets vérifiant les propriétés de A et de B, et la confiance, qui est le quotient de ce support et du nombre d'objets vérifiant les propriétés de A, c'est-à-dire du support de A. On peut aisément transposer ce formalisme de la fouille de données à la classification d'un corpus bibliographique, en prenant pour propriétés les classes et pour objets les mots-clés. Ainsi la figure 2 présente un corpus de ce type caractérisé par 1000 mots-clés et associé à deux ensembles de classes A et B tels que 100 mots-clés caractérisent les classes de A, 300 celles de B, et 80 celles de A et de B, la règle A→B a un support de 80, une confiance de 0,8. Quand la confiance est 1, la règle est toujours vérifiée, on l'appelle alors « règle exacte », elle correspond à la règle d'implication (souvent notée A⇒B) de la logique mathématique, qui n'admet aucun contre-exemple. Ici les 20 mots-clés qui caractérisent les classes de A sans caractériser celles de B sont les « contre-exemples » de la règle A→B, et leur présence fait que la règle est appelée « règle approximative ».

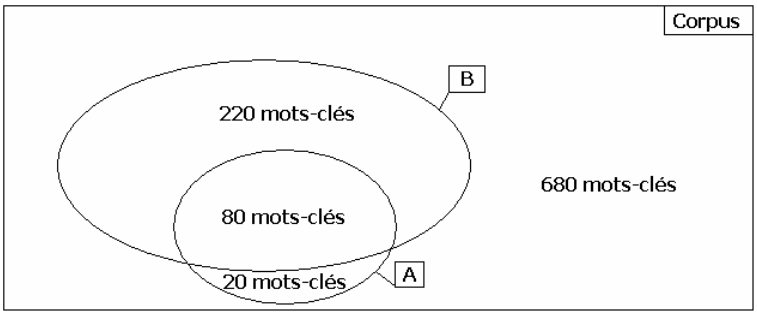

FIG. 2 – *Illustration de la règle A→B.*

On se placera dans le cas où A et B sont deux classes et, pour faciliter l'interprétation, on les identifiera respectivement aux ensembles A' et B' (Guigues et Duquenne, 1986) des mots-clés les caractérisant, en gardant toutefois la notation AB habituelle en fouille de données pour l'ensemble des mots-clés communs à A et B. Ainsi la règle A→B pourra être interprétée comme une inclusion approximative de la classe A dans la classe B. Dans le cas où l'appartenance d'un mot-clé à une règle est de type certain, c'est-à-dire quand les valeurs d'appartenance sont binaires (oui est codé par 1 et non par 0), les algorithmes de règles d'association (Pasquier 2000) permettent d'obtenir très rapidement les règles d'association dont le support et la confiance dépassent des seuils fixés à l'avance.

Quand ces valeurs ne sont plus binaires, c'est-à-dire quand l'appartenance est un nombre variant entre 0 et 1, les règles d'association habituelles, basées sur des effectifs, ne sont plus utilisables. On les remplace alors par des règles d'association floues (Cadot et Napoli, 2004). Du point de vue pratique, on redéfinit l'appartenance aux classes et le support d'une classe de la façon suivante : pour chaque mot-clé, sa valeur d'appartenance à la classe qu'il caractérise correspond à son poids dans la classe. Ainsi, si le mot-clé **i** a le poids $\mathbf{a_i}$ pour la

classe A, et $\mathbf{b_i}$ pour la classe B, sa valeur d'appartenance aux deux classes (noté AB) est la plus petite des deux valeurs, min($\mathbf{a_i}$, $\mathbf{b_i}$). Si les valeurs d'appartenance de tous les mots-clés à une classe ne dépassent pas 0,5, le support de la classe est nul, sinon il est égal à la somme des valeurs d'appartenance. Le tableau 1 décrit ces notions d'appartenance à partir d'un corpus contenant 7 mots-clés contribuant à deux classes A et B correspondant à une classification de type binaire (tableau 1a), et à une classification de type floue (ex : K-means axiales) (tableau 1b).

Mots-clés \ Classes	A	B	AB
MC1	1	0	0
MC2	0	1	0
MC3	0	0	0
MC4	1	1	1
MC5	1	1	1
MC6	0	1	0
MC7	1	1	1
Total	4	5	3
Support de A→B	3		
Confiance de A→B	3/4		

TAB. 1a – *Règle d'association classique*

Mots-clés \ Classes	A	B	AB
MC1	0,6	0,3	0,3
MC2	0,2	0,9	0,2
MC3	0,4	0,9	0,4
MC4	0,9	0,5	0,5
MC5	1	1	1
MC6	0,1	0,9	0,1
MC7	0,8	0,5	0,5
Total	4	5	3
Support de A→B	3		
Confiance de A→B	3/4		

TAB. 1b – *Règle d'association floue*

TAB. 1 – *Exemple de définitions de la règle A→B sur un corpus caractérisé par 7 mots-clés associés à 2 classes A et B : valeurs d'appartenance des 7 mots-clés aux classes A, B et aux deux classes A et B (AB), support et confiance de la règle A→B.*

La confiance de la règle d'association floue est définie comme celle de la règle d'association classique, une fois les supports calculés de la façon indiquée ci-dessus. Avec ces définitions, dans le cas de valeurs binaires, le support et la confiance de la règle d'association floue sont égaux à ceux de la règle d'association classique, et on peut reprendre les algorithmes classiques d'extraction de règles d'association moyennant quelques transformations mineures. L'interprétation d'une règle floue est la même que celle d'une règle classique, c'est une inclusion approximative entre deux classes. Dans le tableau 1a, une grande partie des mot-clés appartenant à A appartiennent aussi à B, la règle d'association classique A→B a donc une confiance élevée (3/4). Dans le tableau 1b, une grande partie des poids des mots-clés de A sont inférieures à leurs poids pour B, la règle d'association floue A→B a la même confiance élevée (3/4).

3 Résultats

Dans cette partie, nous analysons successivement les résultats obtenus avec les trois protocoles puis nous les comparons et discutons.

3.1 Comparaison basée sur les intitulés de classes

Lorsque nous augmentons le nombre de classes, nous constatons non seulement l'apparition de nouvelles classes mais également la disparition de certaines d'entre elles. Si

des intitulés de classes sont toujours présents (Mur soutènement, Stabilité versant, Compression triaxiale…) d'autres peuvent apparaître (Distribution contrainte et Module cisaillement pour Ndoc30), disparaître (Fond marin pour Ndoc50) voire disparaître pour réapparaître (Déformation sous contrainte).

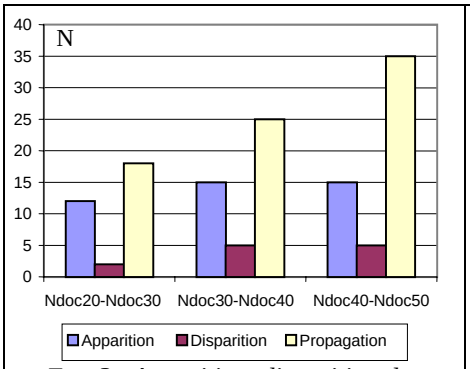

FIG. 3 - *Apparition, disparition des intitulés de classes (N= Nombre de classes)*

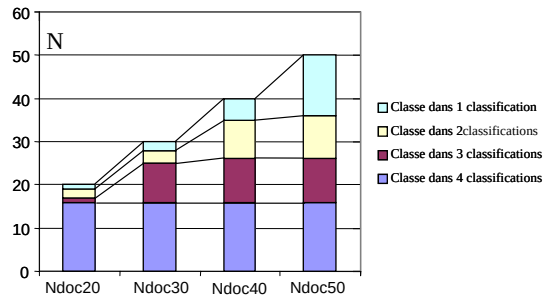

FIG. 4 - *Evolution des classes en fonction de leur intitulé (N= Nombre de classes)*

Pour l'ensemble de ces quatre classifications nous avons 60 intitulés différents dont 16 sont toujours présents. La figure 3 présente le nombre de classes qui apparaissent, disparaissent ou se propagent à chaque changement de classification. Le taux de disparition entre deux classifications est faible pour tous les passages. L'apparition de nouveaux intitulés ainsi que les disparitions se stabilisent à partir du passage Ndoc30-Ndoc40, tandis que le taux de propagation croit de passage en passage. La figure 4 montre l'évolution des intitulés pour chaque classification exprimée en nombre de classes apparaissant n fois à chaque étape. Ces proportions ont tendance à s'égaliser quand le nombre de classes augmente. Dès la première classification nous observons un nombre relativement élevé (16) de classes qui vont se maintenir dans toutes les classifications. Les thèmes présents dans 3 classifications n'apparaissent pas immédiatement. Pour les thèmes présents 2 fois, les classifications semblent liées deux à deux : ainsi Ndoc20 et Ndoc30 ont respectivement 13 à 15% de leurs classes communes à deux classifications alors que Ndoc40 et Ndoc50 en ont toutes deux 20%. Les thèmes isolés (n'apparaissant qu'une fois) sont répartis de façon homogène dans les trois premières classifications et augmentent brutalement pour Ndoc50. Ces deux figures montrent une apparente stabilité des classes avec un faible taux de disparition, et des classes qui se maintiennent au fur et à mesure de leur apparition.

3.2 Comparaison basée sur les documents communs

Le dépouillement des tableaux croisés sur les documents étant lourd, nous avons concentré notre analyse avec cette méthode sur les relations entre les classes de deux classifications successives (nombre de classes croissant). Le tableau 2 présente un extrait de tableau croisé entre les deux classifications Ndoc20 et Ndoc30.

L'analyse de l'ensemble de ces tableaux croisés se concrétise en la réalisation d'un graphe de relation inter-classes dont un extrait est présenté sur la figure 5. Ce graphe permet

de suivre l'évolution de chaque classe et sa « descendance ». Il comprend 155 relations réparties comme suit : 33 relations de Ndoc20 vers Ndoc30, 51 de Ndoc30 vers Ndoc40 et 71 de Ndoc40 vers Ndoc50.

Ndoc20g2x Ndoc30g5 XXXX= 1 0,75 <=XXX< 1 0,5 <=XX<0,75 ; 0,25<=X<0,5 0,1<=x<0,25	Inélasticité	Liquéfaction	Karst	Gestion déchet	Relation contrainte déformation	Résistance cisaillement	Compaction	Consolidation	Essai sol	Résistance compression	**Eau souterraine**	Séisme	Pieu	Compression triaxiale	Pipeline	Déformation sous contrainte	Stabilité versant	Mécanique rupture	**Porosité**
Stabilité versant																	XX		
Inelasticité	x			x						x								x	x
Eau souterraine											XXX								
Teneur eau						X													x
Consolidation							x	X											
Karst			XXX																
Séisme		x										XX							
Déformation sous contrainte														x		X			
Compression triaxiale														X		x			
Pieu													X						

TAB 2 - *Extrait du tableau croisé entre les classifications Ndoc20 et Ndoc30.*

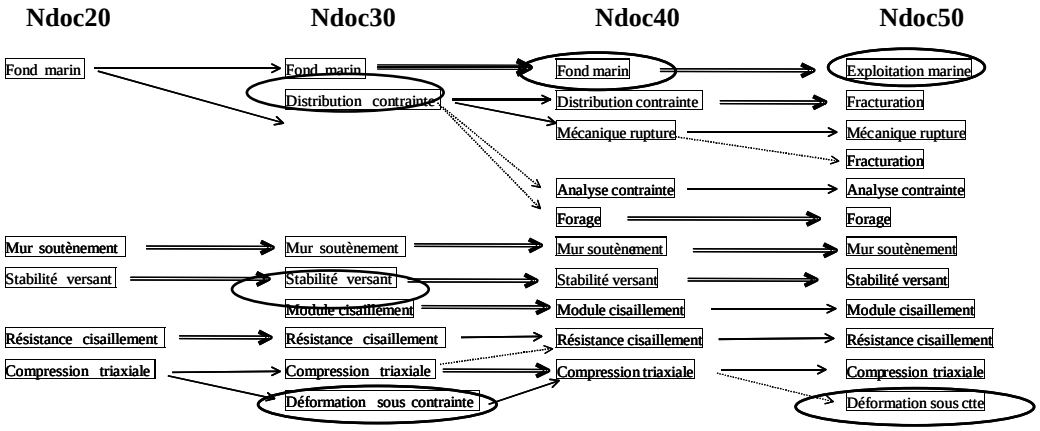

FIG 5 - *Extrait du graphe obtenu à partir des tableaux croisés sur les documents (L'épaisseur de la flèche est fonction de l'indice défini ci-dessus)*

Partant du tableau 2 ou du graphe (figure 5), nous observons différents types d'évolution :

- les classes qui se maintiennent en conservant leur intitulé (ex : Eau souterraine) ou en changeant d'intitulé (ex : Fond marin),
- les classes qui disparaissent par « saupoudrage » des documents dans un grand nombre de classes,
- les classes qui « éclatent » en plusieurs classes (ex : Inélasticité, Distribution contrainte de Ndoc30 à Ndoc40),

- Les classes qui fusionnent (ex : Compression triaxiale et déformation sous contrainte de Ndoc30),
- les classes qui apparaissent (ex : Porosité, Distribution contrainte, ou Module cisaillement dans Ndoc30),
- les classes qui disparaissent pour réapparaître dans d'autres classifications (ex : Déformation sous contrainte).

Les comportements les plus courants sont l'éclatement et la fusion des classes. Nous détaillons ci-dessous les résultats obtenus par éclatement.

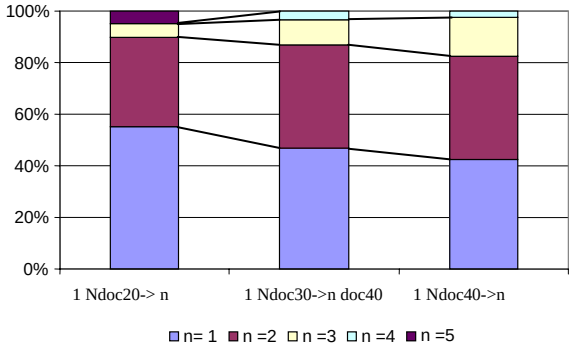

FIG. 6 - *Eclatement des classes*
(1 classe de départ vers n classes d'arrivée)

La figure 6 donne la proportion de classes (en pourcentage du nombre de classes de la classification de départ) qui donnent naissance à n classes (n pouvant aller jusqu'à 5) dans la classification supérieure.

Le nombre de classes stables entre les différents passages (n=1) décroît de 60 à 40%, ce qui est plus faible que les informations obtenues avec la première méthode, où 80% ces classes se propageaient d'une classification à l'autre (figure 3). Cette différence s'explique par le comportement d'une classe comme « Fond marin » qui, entre n=20 et n=30, se scinde en 2 classes dont une conserve l'intitulé initial (figure 5).

Pour l'ensemble des passages entre classifications, 95% des classes présentent des relations avec n variant de 1 à 3, c'est-à-dire qu'une classe est le plus souvent en relation avec au plus 3 classes de la classification supérieure ou inférieure.

3.3 Comparaison basée sur les règles d'association floues

L'extraction des règles d'association floues s'est faite selon la méthode habituelle pour les règles d'association classiques, consistant à prendre des valeurs supérieures à des seuils choisis a priori par l'utilisateur. Dans notre étude les mots clés dont les poids sont inférieurs à 0,2 sont éliminés. Pour les règles le seuil de support est 10 et le seuil de confiance 0,37.

Les règles d'association floues sont établies en croisant toutes les classifications deux à deux et en appliquant une comparaison orientée (par exemple Ndoc20 vers Ndoc40 et Ndoc30 vers Ndoc20). Nous obtenons 409 règles dont 99 concernent des classes appartenant à deux classifications immédiatement voisines et orientées de la plus petite vers la plus

grande. Nous pouvons donc comparer ces 99 règles aux 155 relations du graphe obtenues par les documents communs.

Les règles d'association floues sont représentées sous forme d'un graphe (fig. 7) en utilisant le logiciel Graphviz (interfacé par Hubert et Racodon 2004) où l'épaisseur d'un lien représente l'indice de confiance de la règle, selon les quatre intervalles suivant : 0,37-0,4, 0,4-0,6, 0,6-0,8, et >0,8.

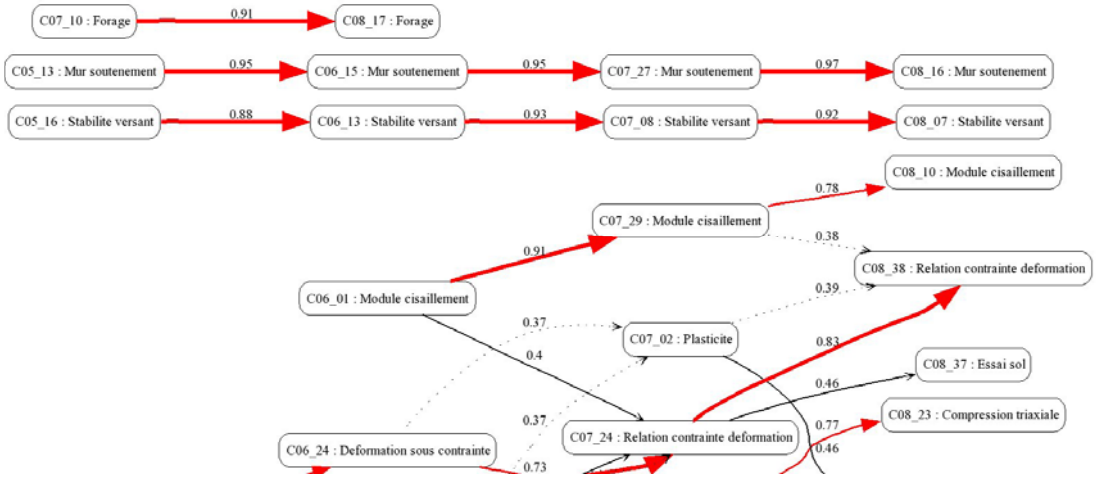

FIG 7 - *Extrait du graphe obtenu à partir des règles d'association floues (C05=Ndoc20, C06=Ndoc30, C07=Ndoc40, C08=Ndoc50)*

Comme pour la méthode précédente, la figure 8 décrit la proportion de classes (en pourcentage du nombre de classes de la classification de départ) qui donnent naissance à n classes dans la classification supérieure. Dans ce dernier cas, n varie de 1 à 3.

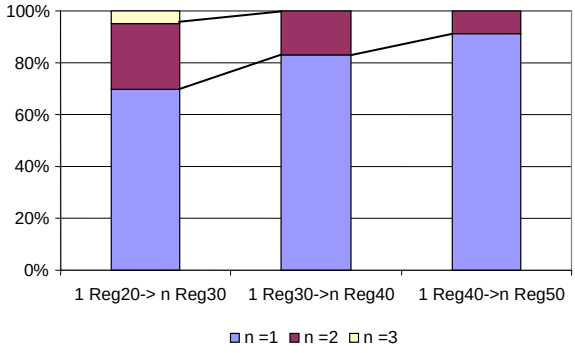

FIG. 8 - *Eclatement des classes (1 classe de départ vers n classes d'arrivée)*

Le nombre de classes stables entre les différents passages (n=1) croît de 65 à 90%, ces valeurs très élevées s'explique par le fait que ce graphe est un sous-ensemble du graphe des relations obtenues à partir des documents communs.

Une étude approfondie des graphes a permis de détecter parmi les 99 relations obtenues avec les règles allant d'une classification à la classification immédiatement supérieure, 95 règles qui sont également des relations établies à partir des documents communs.

3.4 Comparaison entre relations et règles

Pour les 95 règles communes, la figure 9 compare la confiance (C.) des règles d'association floues avec l'indice de force (I.F.) des relations correspondantes calculées à partir des documents communs.

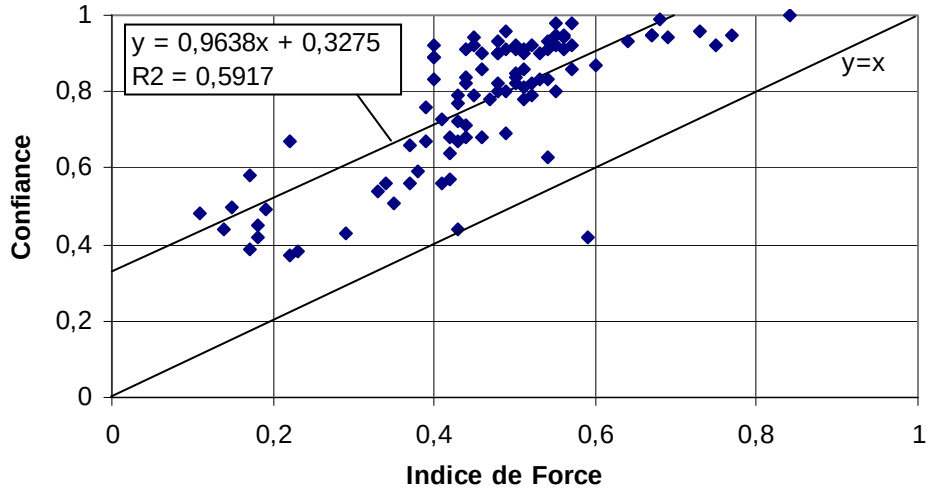

FIG. 9 – *Relation entre confiance et indice de force.*

La figure 9 montre que les deux indices sont bien corrélés et justifie le fait que l'on puisse remplacer l'indice de force par la confiance des règles d'association floues.

4 Discussion et conclusion

La stabilité apparente des classes que nous constatons en étudiant uniquement les intitulés des classes est relativisée quand nous considérons l'analyse des documents communs et des règles d'associations floues. Les règles d'association floues permettent d'analyser de façon complète les relations orientées entre toutes les classifications (ici plus de 400 règles). L'analyse des tableaux croisés en documents communs se révélant plus fastidieuse, nous l'avons limitée à la comparaison 2 à 2 des classifications ayant un nombre croissant de classes. En considérant uniquement les règles qui peuvent être comparées aux résultats d'analyse sur les documents communs, nous constatons que pratiquement la totalité de

celles-ci (96%) sont validées à partir des tableaux croisés. De plus l'Indice de Force des relations croît dans le même sens que la confiance des règles correspondantes.

Cette expérimentation permet de valider, sur la comparaison des classifications de précision croissante (nombre de classes croissant), la convergence entre les informations apportées par les règles d'associations floues et les relations obtenues à partir des tableaux croisés des documents. Ces règles sont obtenues de façon automatique, et sont appliquées sur l'ensemble des croisements possibles entre classifications. Elles peuvent donc avantageusement remplacer l'analyse des tableaux croisés. Avec les règles d'association floues, nous disposons donc d'une méthode fiable pour comparer des classifications uniquement à partir des descripteurs. Ceci nous permet de comparer des classifications réalisées sur des corpus différents situés dans un espace de description relativement stable.

Il est donc envisageable d'aborder le problème du suivi de l'évolution dans le temps par cette méthode ; en réalisant une classification à un temps T, puis une nouvelle classification avec le corpus correspondant à T+ΔT, et enfin en définissant à partir des règles d'association les classes ayant évolué. Une autre approche serait de réaliser une classification pour ΔT puis, grâce aux règles d'association floues, d'analyser les relations entre les classes correspondant au corpus à ΔT et celles correspondant au corpus à T (inclusions, nouvelles classes…). Il serait alors possible, dans le cas d'inclusion, d'affecter les nouveaux documents aux classes T, et dans le cas de nouvelles classes de les rajouter à la classification T.

Remerciements : Ce travail a été entrepris dans le cadre du projet Ingénierie des Langues, du Document et de l'Information Scientifique, Technique et Culturelle (ILD&ISTC) du pôle Intelligence logicielle du Contrat Plan Etat – Région en Lorraine. Nous remercions Jean-Paul TISOT directeur de l'Ecole Nationale Supérieure de Géologie pour son aide et son analyse experte du corpus et des classifications ainsi que Marie HUBERT et David RACODON pour leurs développements informatiques.

Références

Cadot M., Napoli A. (2004), Règles d'association et codage flou des données, 11èmes Rencontres de la Société Francophone de Classification, SFC 2004 (Bordeaux), pp. 130-133.

Diday E. (1972), Optimisation en classification automatique et reconnaissance des formes. Revue française d'Automatique, Informatique et Recherche Opérationnelle, vol. 3, pp. 61-95.

Forgy E. W. (1965), Cluster analysis of multivariate data : efficiency versus interpretability of classifications, Biometrics, vol. 21, n° 3, p. 768.

Guigues J.L. et Duquenne V. (1986), Familles minimales d'implications informatives résultant d'un tableau de données binaires, Math. Sci. Hum., n°95, pp. 5-18.

Graphviz, www.research.att.com/sw/tools/graphviz.

Han J., Kamber M. (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco.

Hubert M., Racodon D. (2004), Rapport projet 2A, Data Mining, stage d'initiation à la recherche ESIAL Nancy.

Kodratoff Y. (2001), Rating the Interest of Rules Induced from Data and within texts, 12th IEEE -International Conference on Database and Expert Systems Aplications-Dexa 2001, Munich, sept 2001, pp. 265-269.

Lamirel J.C., François C., Shehabi S. AL, Hoffman M. (2004), New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping, Scientometrics, Vol.60, N°3, pp. 445-462.

Lelu A. (1993), Modèles neuronaux pour l'analyse de données documentaires et textuelles. Paris, Thèse de l'Université de Paris VI, 238 pages.

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.

McQueen J.B. (1967), Some methods of classification and analysis of multivariate observations, L. Le Cam and J. Neyman (Eds.), Proc. 5th Berkeley Symposium in Mathematics, Statistics and Probability, vol 1., pp. 281-297, Univ. of California, Berkeley, USA, 1967.

Pasquier N., (2000), Data Mining : Algorithmes d'Extraction et de Réduction des Règles d'Association dans les Bases de Données, Thèse, Université de Clermont-Ferrand II.

Polanco X., Francois C. (2000), Data Clustering and Cluster Mapping or Visualization in Text Processing and Mining, in: Dynamism and Stability in Knowledge Organization, proceceedings of the Sixth international ISKO conference, 10-13 July 2000, Toronto, Canada. Edited by Clare Beghtol, Lynne C. Howarth, Nancy J. Williamson : Ergon Verlag, 2000, pp 359 - 365.

Polanco X., François C., Royauté J., Besagni D (2001), STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology, 8th International Conference on Scientometrics and Informetrics, July 16-20 2001, Sydney, Australia, Proceedings Vol 2, pp. 871 – 873.

Youness G., Saporta G. (2004), Une méthodologie pour la comparaison de partitions, Revue Statistique Appliquée, Vol. 52, n° 1, pp. 97-120.

Summary

Our work is based on the analysis of a bibliographical corpus in the domain of geotechnics using maps realised with the Stanalyst® platform. This tool integrates a non-hierarchical clustering algorithm (axial K-means) the results of which depend on the required number of clusters. This instability makes any comparison between clusterings difficult, and introduces a doubt as to the right number of clusters necessary to represent correctly a domain. We compare the results of a cluster analysis according to 3 protocols; (1) analysis of the cluster names; (2) relationships between clusters based on their common members; (3) fuzzy association rules. As the graphs we obtained show remarkable similarities, we express a preference for the fuzzy association rules for they are extracted automatically from the cluster descriptions and not from their members. This allows us to analyse clusterings from different corpora.