L'analyse relationnelle pour la fouille de grandes bases de données.

Hamid Benhadda*, François Marcotorchino*

*160, boulevard de Valmy – BP 82
92704 Colombes Cedex
{ hamid.benhadda,jeanfrancois.marcotorchino }@fr.thalesgroup.com

Résumé. Dans cet article nous montrerons, brièvement, les possibilités offertes par la théorie de l'analyse relationnelle, initiée dans les années 1980 à IBM-Corp. Nous nous concentrerons sur les avancées théoriques et méthodologiques obtenues grâce à cette théorie pour fusionner l'information et pour traiter et analyser de grandes quantités de données qu'elles soient de type structuré ou non structuré. Nous aborderons brièvement la théorie de la similarité régularisée, théorie basée sur l'analyse relationnelle et la généralisant mais plus récente. Nous montrerons aussi des formules de transfert permettant d'exprimer des problèmes combinatoires bien connus sous forme de fonctions économiques linéaires appropriées pour différents type de problématique (tels que des problèmes de classification automatique ou des problèmes d'association,). Ceci en plus de la complexité linéaire O(N) de l'algorithmique sous jacente qui permet à cette approche d'être tout à fait convenable pour différentes applications réelles.

1 Introduction

De nos jours plus encore qu'à d'autres époques, les progrès techniques et scientifiques d'une part et les faibles coûts de stockage d'autre part poussent l'homme à rassembler et conserver des quantités de plus en plus grandes de données. Cette accumulation de données, est amplement justifiée par le fait qu'à notre époque, la possession et l'exploitation du maximum d'information confèrent à ceux qui les maîtrisent un avantage concurrentiel majeur.

Il devient donc nécessaire d'avoir à sa disposition des outils d'analyse et d'exploitation de ces données, afin d'en extraire une information à valeur ajoutée qui pourra être utilisée par la suite pour faire de nouveaux progrès dans les domaines particuliers relatifs à ces données.

Ceci ne pourra se réaliser que si les outils concernés respectent la structure des données qu'on leur confie. En particulier, ces outils doivent permettre de manipuler, de combiner et de structurer les variables et attributs d'analyse, en les considérant comme des entités propres et séparées et non comme un magma global qu'on considèrera comme un tout ou en adaptant les données aux méthodes préexistantes, en violant leur nature pour satisfaire les exigences de ces méthodes.

Afin de respecter les exigences qui viennent d'être citées, nous allons parler, dans cet article, d'une part, de l'analyse relationnelle et de ses extensions et d'autre part de la similarité régularisée, théorie qui a été co-développée par les auteurs et qui généralise la théorie de l'analyse relationnelle.

2 L'analyse relationnelle

L'analyse relationnelle est une technique d'analyse des données à vaste champ applicatif. Elle a été initiée et développée par F. Marcotorchino et P. Michaud (1978) au Centre Européen de Mathématiques Appliqués (ECAM) à IBM. Cette technique, encore assez méconnue, utilise le concept de « comparaisons par paires » dont l'apparition dans la littérature statistique se fait vers la fin des années trente dans les travaux de M.G. Kendall et B. Smith[KB40], bien que le concept dont ce sont inspirés les auteurs précédents, date des travaux du marquis de Condorcet en 1875, autour de la théorie des votes.

De façon générale, l'analyse relationnelle permet de modéliser et de résoudre des problèmes dont la formulation générale peut s'énoncer : ***« Rechercher une relation particulière S qui s'ajuste « au mieux » à une (ou plusieurs) relation(s) quelconque(s) donnée(s) R. »***

Cette théorie a été mise au point pour résoudre deux catégories de problèmes majeurs que les utilisateurs rencontrent souvent lorsqu'ils démarrent des processus avancés de traitement de grandes quantités de données : « capacité de traiter de grands tableaux de données » et « robustesse des processus d'analyse » quelle que soient leur nature. Un des principes latents que l'on rencontre, dans ce contexte, est connu sous la nom de « principe de décomposition ». Ce principe consiste à subdiviser la population hétérogène globale en sous groupes (sans fixer a priori, comme c'est le cas des autres techniques, ni leur nombre ni la taille de leur population) dans le but d'obtenir des group homogènes en terme de similarité. Ce gain en homogénéité permet l'application, ensuite à l'intérieur des groupes obtenus, d'autres techniques telles que : la régression, le scoring, les arbres de décision, …etc. Ce qui donne de meilleurs résultats en terme de robustesse et de qualité d'analyse.

L'analyse relationnelle, comme toute technique de classification, prend comme point d'entrée une matrice rectangulaire, représentant les données à classifier. Les lignes de la matrice représentent les individus à classifier et les colonnes les variables mesurées sur ces individus. L'intersection d'une ligne et d'une colonne du tableau de données est la valeur prise par la variable colonne sur l'individu ligne. Le point commun à tous les problèmes traités par l'approche relationnelle, est la prise en compte des données de départ dans une matrice de comparaisons par paires C (appelée matrice de Condorcet) de terme général $c_{ii'}$ représentant l'accord ou similarité entre l'individu i et l'individu i' relativement à l'ensemble des variables mesurées sur la population totale.

Dans la suite du document nous supposerons que n est le nombre d'individus de la population étudiée et que m est le nombre de variables mesurées sur ces individus.

2.1 Quelques avantages spécifiques

Dans la majorité des logiciels de data mining existants, les techniques de classification qu'ils utilisent lorsqu'ils traitent de grandes quantités de données imposent :

- De fixer a priori le nombre de classes à trouver dans la population d'origine et
- De faire de l'échantillonnage, à cause des limitations en terme de quantités d'individus que les algorithmes peuvent gérer en des temps raisonnables.

Dans la théorie relationnelle, d'une part, il n'y a pas de fixation arbitraire et a priori du nombre de classes et d'autre part, on n'utilise pas d'échantillonnage pour traiter les données.

Ces deux avantages, ne sont que quelques uns parmi bien d'autres, comme nous le verrons dans la suite de l'article.

2.2 La méthodologie

Le modèle mathématique à la base de la théorie relationnelle est loin d'être trivial. Le propos de cet article sera de montrer quelques principes simples sous jacents à cette théorie. Un des points importants de cette technique est qu'elle utilise le concept de « comparaisons par paires » dont l'apparition dans la littérature statistique se fait vers la fin des années trente dans les travaux de M.G. Kendall et B. Smith[KB40] et au niveau concept, dès 1785 avec les premières approches du marquis de Condorcet sur la théorie des votes.

2.2.1 Principe de comparaisons par paires

Pour illustrer le « principe de comparaisons par paires », supposons que la population étudiée soit composée de cinq personnes notées {P1, P2, P3, P4, P5} et que la variable mesurées sur ces personnes soit « la nationalité » pouvant prendre les modalités {Française, Espagnole, Anglaise}. Supposons que les deux premières personnes soient de nationalité française, la troisième de nationalité espagnole et les deux dernières de nationalité anglaise. Le principe de « comparaisons par paires » consiste à transformer, dans notre exemple, la variable « nationalité » en une matrice carrée C de dimensions (5×5) de terme général $c_{ii'}$ prenant les valeurs 1 ou 0, selon que les individus i et i' sont de même nationalité ou non, i.e. :

$$c_{ii'} = \begin{cases} 1 & si\ i\ et\ i'\ sont\ de\ même\ nationalité \\ 0 & autrement \end{cases}$$

le terme $c_{ii'}$, peut être interprété comme une mesure de similarité entre les individus i et i'. En effet, si deux individus ont la même nationalité, ils sont considérés comme étant similaires par rapport à cette variable. La matrice obtenue pour notre exemple est :

Données		Représentation relationnelle C				
Nationalité	Individus	P1	P2	P3	P4	P5
Français	P1	1	1	0	0	0
Français	P2	1	1	0	0	0
Espagnol	P3	0	0	1	0	0
Anglais	P4	0	0	0	1	1
Anglais	P5	0	0	0	1	1

TAB. 1 – *Principe de comparaisons par paires.*

Il existe une correspondance entre la représentation linéaire de la variable nationalité (sous forme de vecteur) et sa représentation relationnelle C, mais cette correspondance n'est pas bi-univoque car la représentation relationnelle est plus générale que la représenta-

tion vectorielle. En effet, contrairement à la représentation vectorielle, la représentation relationnelle peut gérer l'appartenance à plusieurs catégories. Si par exemple la personne P3 possède les trois nationalités citées ci-dessus, aucune représentation vectorielle de cette information n'est possible, par contre il suffirait de mettre un 1 dans la ligne et la colonne correspondant à P3, ce qui donnerait le tableau suivant :

	Représentation relationnelle				
	C				
Individus	P1	P2	P3	P4	P5
P1	1	1	1	0	0
P2	1	1	1	0	0
P3	1	1	1	1	1
P4	0	0	1	1	1
P5	0	0	1	1	1

TAB. 2 – Gestion relationnelle des multi-catégories.

La représentation relationnelle permet aussi, contrairement à la représentation vectorielle, de traiter les boucles lorsqu'il s'agit de traiter des relations d'ordre. Par exemple, il est possible de représenter la situation « P1>P2>P3>P4>P5>P1 » où le signe « > » signifie « est préféré à ».

2.2.2 Classification par la méthodologie relationnelle

Pour classifier une population formée de n individus $(O_1, O_2, \ldots, O_n)$ décrits par m variables $(V^1, V^2, \ldots, V^m)$, on commence par transformer chaque matrice V^k en une matrice relationnelle C^k de terme général $c_{ii'}^k$ représentant la similarité entre les deux individus par rapport à la variable V^k . Une fois toutes les matrices C^k obtenues, on construit la matrice relationnelle globale C de terme général $c_{ii'}$ comme la somme des mesures de similarité des deux individus sur l 'ensemble des variables :

$$c_{ii'} = \sum_{k=1}^{m} c_{ii'}^k$$

Une propriété importante de toute mesure de similarité est «l'auto similarité maximale». Cette propriété stipule que la similarité d'un individu avec lui-même est toujours supérieure ou égale à sa similarité avec n'importe quel autre individu i.e. pour tout individu i :

$$c_{ii'} \leq c_{ii} \qquad \forall i'$$

ou de façon plus générale :

$$c_{ii'} \leq Min(c_{ii}, c_{i'i'}) \qquad \forall i, i'$$

On en déduit que $Min(c_{ii}, c_{i'i'})$ est la «similarité maximum possible»entre deux individus i et i' donnés. A partir de la similarité et de la « similarité maximum possible » entre

ces deux individus , on définit leur dissimilarité comme le complément de leur similarité à leur « similarité maximum possible » :

$$\overline{c}_{ii'} = Min(c_{ii}, c_{i'i'}) - c_{ii'}$$

ces deux individus seront, a priori, dans la même classe de la partition finale dès lors que leur similarité sera supérieure à leur dissimilarité :

$$c_{ii'} > \overline{c}_{ii'}$$

La partition finale recherchée sera représentée par une matrice carrée binaire X dont le terme général $x_{ii'}$ est défini par la relation :

$$x_{ii'} = \begin{cases} 1 & si \ \ i \ et \ i' \ sont \ dans \ la \ même \ classe \ de \ la \ partition \ finale \\ 0 & dans \ le \ cas \ contraire \end{cases}$$

Cette partition étant une relation d'équivalence, elle doit respecter les contraintes de :

- Réflexivité : un individu est dans la même classe que lui-même

$$x_{ii} = 1$$

- Symétrie : si l'individu i est dans la classe de l'individu i', alors i' est dans la classe de i

$$x_{ii'} = 1 \quad \Rightarrow \quad x_{i'i} = 1$$

- Transitivité : si l'individu i est dans la même classe que l'individu i' et que l'individu i' est dans la même classe que l'individu i'', alors i est dans la même classe que i''

$$(x_{ii'} = 1 \ et \ x_{i'i''} = 1) \quad \Rightarrow \quad x_{ii''} = 1$$

La partition X sera obtenue après maximisation du critère de Condorcet $C(X)$ suivant :

$$C(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} \left(c_{ii'} x_{ii'} + \overline{c}_{ii'} \overline{x}_{ii'} \right)$$

avec : $\overline{x}_{ii'} = 1 - x_{ii}$

La formulation mathématique du problème relationnel à résoudre consiste à trouver la partition optimale X telle que :

$$\underset{X}{Max} \left[\sum_{i=1}^{n} \sum_{i'=1}^{n} \left(c_{ii'} x_{ii'} + \overline{c}_{ii'} \overline{x}_{ii'} \right) \right]$$

X *vérifiant* :

$$\begin{cases} x_{ii} = 1 & \forall i & (réflexivité) \\ x_{ii'} = x_{i'i} & \forall i, i' & (Symétrie) \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 & \forall i, i', i'' & (transitivité) \\ x_{ii'} \in \{0,1\} & & (binarité) \end{cases}$$

On peut montrer facilement que maximiser le critère de Condorcet revient à maximiser le critère $C'(X)$ suivant, sous les mêmes contraintes de réflexivité, de symétrie et de transitivité donnée ci-dessus :

$$C'(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} \left(c_{ii'} - \frac{Min(c_{ii}, c_{i'i'})}{2} \right) x_{ii'}$$

La solution exacte de ce problème s'obtient par programmation linéaire dans le cas où le nombre d'individus à classifier serait relativement petit, mais dans la pratique on a recours à une heuristique qui permet l'obtention d'une solution approchée.

2.3 Etapes de la première heuristique relationnelle

Comme nous l'avons indiqué au paragraphe (2.1), dès que le nombre de données devient important, on a recours à des heuristiques pour chercher la solution la plus proche possible de la solution exacte (celle que l'on obtiendrait par programmation linéaire). Nous allons donner ci-dessous la description de la première heuristique qui a été mise en œuvre au tout début de l'utilisation de la méthodologie relationnelle. Une seconde heuristique, plus récente et qui dépasse le cadre de cet article, est mise en œuvre actuellement dans les algorithmes de classification automatiques relationnelle.

2.3.1 Etape 1 : Initialisation

L'initialisation consiste à partir de la population de départ et à former les classes au fur et à mesure selon les étapes suivantes :

- Prendre un individu quelconque et le mettre dans la première classe
- Prendre un deuxième individu, si son accord avec la classe précédente constituée d'un seul individu est supérieur à son désaccord avec cette même classe, alors mettre les deux individus dans la même classe, sinon créer une nouvelle classe et y mettre ce second individu
- Prendre un troisième individu, calculer son accord avec les deux classes existantes et le mettre dans la classe avec laquelle il a le meilleur accord, sinon créer une nouvelle classe et y mettre cet individu
- Continuer ainsi jusqu'à ce que chaque individu de la classe soit affecté à une classe. .

2.3.2 Etape 2 : Réunion de deux classes

A l'issue de l'étape d'initialisation on se trouve avec un certain nombre de classes. Il s'agit ensuite de prendre les classes les unes après les autres, calculer pour chaque classe considérée son accord avec les autres classes et la réunir avec la classe avec laquelle l'accord est le plus grand (si cet accord est supérieur à leur désaccord). Ceci doit être réalisé tant qu'il y a une possibilité d'améliorer le critère $C'(X)$.

2.3.3 Etape 3 : transfert d'un individu d'une classe à une autre

Quand aucune réunion n'est plus possible, on prend les individus de chaque classe un par un, on calculera l'accord de chaque individu avec chaque classe autre que la sienne propre. Si un individu a un accord meilleur avec une autre classe que la sienne et si le désaccord avec cette nouvelle classe est inférieur à l'accord alors cet individu sera transféré de sa classe à la nouvelle classe, avec laquelle il a l'accord maximum. Ceci sera poursuivi jusqu'à ce qu'il n'ait plus de possibilité d'amélioration du critère.

2.3.4 Etape 4 : réunion de deux classes

Quand, aucun transfert des individus d'une classe à une autre n'est plus possible, on retourne à l'étape 2, pour voir s'il n'est pas possible d'améliorer le critère de Condorcet en réunissant d'autres classes. Ces quatre étapes seront appliquées, jusqu'à ce qu'il n'y ait plus d'amélioration du critère.

2.4 Indicateurs mesurant la qualité de la partition obtenue

Plusieurs indicateurs sont calculés pour mesurer la qualité de la partition finale obtenue. Tous ces indicateurs sont compris entre 0 et 1. A cet effet, nous allons donner quelques définitions qui seront utilisées dans la suite de ce document.

On posera :

$$\kappa = Nombre \quad de\ classes \quad obtenues$$

$$A_{cc'} = \sum_{i \in C} \sum_{i' \in C'} c_{ii'}$$

$$\overline{A}_{CC'} = \sum_{i \in C} \sum_{i' \in C'} \overline{c}_{ii'}$$

$$AM_{cc'} = \sum_{i \in C} \sum_{i' \in C'} Min\left(c_{ii}, c_{ii'}\right)$$

2.4.1 Qualité de la partition obtenue

Cet indicateur mesure la cohérence globale de la partition résultat obtenue :

$$Q = \frac{\sum_{C=1}^{\kappa} A_{CC} + \sum_{C=1}^{\kappa} \sum_{C' \neq C} \overline{A}_{CC'}}{\sum_{C=1}^{\kappa} \sum_{C'=1}^{\kappa} AM_{CC'}}$$

2.4.2 Qualité d'une classe particulière

Cet indicateur mesure l'homogénéité d'une classe C donnée, en prenant en compte à la fois l'homogénéité interne de la classe et ses liaisons avec les autres classes :

$$Q_C = \frac{A_{CC} + 2 \times \sum_{C' \neq C} \overline{A}_{CC'}}{AM_{CC} + 2 \times \sum_{C' \neq C}^{K} AM_{CC'}}$$

2.4.3 Intra d'une classe

Cet indicateur tient compte uniquement de l'homogénéité propre à une classe C donnée :

$$I_C = \frac{\sum_{i \in C} \sum_{i' \in C} c_{ii'}}{\sum_{i \in C} \sum_{i' \in C} Min(c_{ii}, c_{i'i'})}$$

2.4.4 Inter de deux classes

Cet indicateur mesure le lien qu'entretiennent entre elles deux classes différentes données C et C' :

$$I_{CC'} = \frac{\sum_{i \in C} \sum_{i' \in C'} c_{ii'}}{\sum_{i \in C} \sum_{i' \in C'} Min(c_{ii}, c_{i'i'})}$$

2.5 Exemple d'illustration

Supposons que la population étudiée soit composée de sept individus $(O_1, O_2, \ldots, O_7)$ sur lesquels ont été mesurées trois variables qualitatives (V^1, V^2, V^3). Les données étant représentées dans le tableau suivant :

	V^1	V^2	V^3
O_1	1	1	1
O_2	1	1	1
O_3	1	2	2
O_4	2	2	2
O_5	2	2	2
O_6	3	2	3
O_7	3	3	3

TAB. 3 – Données d'origines.

Après transformation des trois variables vectorielles en leurs représentations relationnelles et sommation des ces dernières, on obtient la matrice globale de Condorcet C suivante :

	C						
	O_1	O_2	O_3	O_4	O_5	O_6	O_7
O_1	3	3	1	0	0	0	0
O_2	3	3	1	0	0	0	0
O_3	1	1	3	2	2	1	0
O_4	0	0	2	3	3	1	0
O_5	0	0	2	3	3	1	0
O_6	0	0	1	1	1	3	2
O_7	0	0	0	0	0	2	3

TAB. 4 –Matrice globale de Condorcet.

Comme $c_{ii} = 3$ pour tout individu i, la « similarité maximum possible » entre deux individus quelconques est donc égale à 3. On en déduit que $\bar{c}_{ii'} = 3 - c_{ii'}$, quel que soient les individus i et i'. La partition X obtenue est constituée des trois classes suivantes :

- Classe 1 : O_1, O_2
- Classe 2 : O_3, O_4, O_5
- Classe 3 : O_6, O_7

Cette partition est représentée par la matrice binaire suivante :

	X						
	O_1	O_2	O_3	O_4	O_5	O_6	O_7
O_1	1	1	0	0	0	0	0
O_2	1	1	0	0	0	0	0
O_3	0	0	1	1	1	0	0
O_4	0	0	1	1	1	0	0
O_5	0	0	1	1	1	0	0
O_6	0	0	0	0	0	1	1
O_7	0	0	0	0	0	1	1

TAB. 5 –Matrice binaire représentant la partition obtenue.

La valeur du critère de Condorcet $C(X)$ correspondant est égale à 131, et la qualité de la partition est égale à 0.89.

2.6 Quelques extensions de la théorie relationnelle

Comme nous l'avons signalé dans l'introduction, l'analyse relationnelle a été à l'origine de plusieurs autres théories ou techniques qui sont utilisées dans le domaine de l'analyse des données. Nous allons citer, brièvement, dans les paragraphes qui suivent quelques unes de ces techniques, et détailler un peu plus deux d'entre elles : la similarité régularisé et la linéarisation des critères de contingence.

2.6.1 La théorie des préférences

La théorie relationnelle, dont les origines remontent au marquis de Condorcet à été utilisée en premier par ce dernier pour résoudre le problème du consensus de préférences lors d'un vote. Ce problème a été traité en détail et généralisé ensuite par F . Marcotorchino et P. Michaud (1978) et plus particulièrement, par P. Michaud (1981, 1985).

2.6.2 La sériation

Contrairement à la classification du type Condorcet, qui utilise l'espace des variables pour classifier l'espace des individus, la sériation (ou bi-clustering pour les anglophones) classifie simultanément l'espace des individus et l'espace des variables. L'analyse relationnelle a été utilisée avec sucées dans ce type de problèmes, en particulier dans le domaine de la productique par F. Marcotorchino (1987,1991).

2.6.3 L'analyse factorielle-relationnelle

F. Marcotorchino (1991) a utilisé conjointement la théorie relationnelle et l'analyse factorielle des correspondances pour répondre au problème de fixation du nombre de classes lorsqu'on cherche à classifier une population en utilisant des critères inertiels.

2.7 La similarité régularisée

La similarité régularisée co-développée par les auteurs (1997, 1998) est une théorie qui est venue enrichir et généraliser la théorie de l'analyse relationnelle classique. L'idée fondamentale sur laquelle se base cette théorie est que les variables ont des structures internes non décelables a priori qui confèrent implicitement à certaines d'entre elles des poids plus importants qu'à d'autres dans le processus d'analyse. En effet, il semble évident les variables à deux modalités, par exemple, auront tendances à générer plus de similarités que les variables plus grand nombre de modalités. Par exemple, il est plus difficile pour deux personne, choisies au hasard dans Paris, d'habiter dans le même arrondissement (20 modalités) que d'être du même sexe (2 modalités).

Le principe de la similarité régularisée est de tenir compte, dans le calcul des similarités entre les individus, de ces structures internes afin de compenser (ou rééquilibrer) les influences, trop fortes ou trop faible, induites de façon implicite par ces structures.

A cet effet, on commence par définir, pour chaque variable V^k, une similarité intrinsèque $s^k_{ii'}$ entre les deux individus i et i', on définit ensuite un poids $\pi^k_{ii'}$ traduisant degré de création de similarité induit par de cette variable. Plus ce degré est grand, plus le poids que l'on va lui attribuer sera faible. Par exemple, si l'on veut tenir compte des modalités pour définir $\pi^k_{ii'}$, il suffit de poser : $\pi^k_{ii'} = \frac{1}{p_k}$ où p_k est le nombre de modalités de la variable V^k. On définira ensuite la similarité régularisée $sr^k_{ii'}$ par la relation :

$$sr^k_{ii'} = s^k_{ii'}\left(1-\pi^k\right) = s^k_{ii'}\left(1-\frac{1}{p_k}\right)$$

Dans ce cas on voit bien que plus p_k est grand plus la similarité $sr^k_{ii'}$ est forte. Mais ceci est un cas simpliste, il existe aussi des variantes de la fonction $\pi^k_{ii'}$ reposant sur des considérations statistiques voire pour les plus complexes (structures densitaires), exprimables uniquement via le recours à des notations de la théorie de l'Analyse Relationnelle (représentation matricielle de chaque variable par des graphes de relations binaires) (voir par la suite).

D'ailleurs on peut montrer mathématiquement que la similarité Régularisée globale est la somme (ou la moyenne arithmétique) des similarités régularisées de chaque variable selon la formule suivante, qui intègre en une seule formulation l'agrégation de similarités complexes calculées sur chacune des variables :

$$sr_{ii'} = \frac{1}{m}\sum_{k=1}^{m} sr^k_{ii'} = \frac{1}{m}\sum_{k=1}^{m} s^k_{ii'}\left(1-\pi^k_{ii'}\right)$$

Nous allons donner, dans la suite des exemples de fonction de similarité régularisée plus complexe.

Si nous prenons à titre d'exemple comme indice de similarité « sémantique » sur la variable V^k (cas où la variable est qualitative), un indice de similarité dit de « présence-rareté »[1] par :

$$sr_{ii'}^k = \frac{s_{ii'}^k}{\sum_{i'=1}^{n} s_{ii'}^k} = \begin{cases} \frac{1}{s_{i.}^k} & si\ i\ et\ i'\ sont\ de\ la\ même\ catégorie\ de\ V^k \\ 0 & dans\ le\ cas\ contraire \end{cases}$$

Si par ailleurs on définit une fonction $\pi_{ii'}^k$ de difficulté de « matching », suivant la formule suivante :

$$\pi_{ii'}^k = \frac{\sum_{j=1}^{n}\sum_{j'=1}^{n} s_{jj'}^k}{n^2} = \frac{s_{..}^k}{n^2}$$

Cette fonction représentant dans le cas général la « densité » de « 1 » dans la matrice relationnelle représentant la variable V^k. En particulier si toutes les modalités sont équiréparties dans la population (c'est à dire) chacune regroupe des effectifs d'objets de même taille, la fonction ci-dessus nous redonne le cas déjà décrit précédemment :

$$\pi_{ii'}^k = \frac{1}{p_k}$$

à ce propos, les fonctions de « difficulté de matching » données ci dessus, sont des constantes ne dépendant que de V^k, mais une fonction comme celle qui suit, correspondant à une configuration que nous ne développerons pas, est, elle, bien dépendante de i et i' :

$$\pi_{ii'}^k = \frac{\sum_{i'=1}^{n} s_{ii'}^k + \sum_{i'=1}^{n} s_{ii'}^k}{2n} = \frac{s_{i.}^k + s_{.i'}^k}{2n}$$

On pourrait, aussi utiliser, une « double régularisation » qui consiste à prendre en compte à la fois la « présence-rareté » et la « difficulté de matching ». La similarité s'écrirait donc :

$$sr_{ii'}^k = \frac{s_{ii'}^k}{s_{i.}^k}\left(1 - \frac{s_{..}^k}{n^2}\right)$$

---

1 Un indice de similarité de « présence –rareté » entre deux objets i et j est d'autant plus fort que i et j sont peu nombreux à partager la même valeur de V^k (d'où le concept de « rareté »).

A titre de conclusion, L'approche classique revient à calculer la similarité entre objets de façon longitudinale (ou horizontale) en prenant en compte les vecteur lignes de la matrice de données, tandis que l'approche par similarité régularisée commence par calculer la similarité entre individus, séparément variable par variable (en tenant compte de la sémantique propre à chaque variable) de façon verticale (approche par colonnes), puis somme dans un deuxième temps les similarités calculées variable par variable, en une similarité agrégée globale

On peut trouver plus de détails concernant cette théorie dans H. Benhadda (1998).

2.8 Liaison entre l'approche relationnelle et l'approche contingentielle

Lorsqu'il s'agit d'analyser de très grandes masses de données, une étape, préalable et indispensable, avant l'utilisation des algorithmes d'analyse proprement dits, consiste à pré-traiter les données afin de les nettoyer des erreurs et bruits qui peuvent les affecter. Parmi les techniques statistiques les plus utilisées lors de ces pré-traitements on trouve la recherche des corrélations entre les variables. Cette recherche peut aider, par exemple, à réduire l'espace de description en éliminant de l'analyse une des deux variables jugées très corrélées. Ce qui permettra d'accélérer le processus d'analyse sans perte notable d'information.

Lorsque les variables mesurées sur la population étudiée sont de type qualitatif, il existe une panoplie d'indices d'association inventés dans le but de mesurer les corrélations que ces variables peuvent entretenir entre elles. Dans la pratique courante dans le domaine de l'analyse des données, lorsqu'on se trouve en présence de variables qualitatives (partitions), on a recours à un codage binaire de l'information appelé : « codage disjonctif complet ». Cette forme de codage consiste à diviser une variable qualitative en autant de variables binaires qu'elle possède de catégories (ou modalités). Grâce à ce codage et à l'utilisation des formulations relationnelles, nous montrerons qu'il est possible de linéariser certains critères parmi les plus utilisés en statistiques des contingences. Nous appliquerons, à titre d'exemples, ces linéarisations à trois critères : le critère du Chi-deux, le crière de Belson et celui de Rand.

2.8.1 Propriétés relatives à une seule variable

Si n est le nombre d'individus constituants la population et que la variable V, mesurée sur individus, possède p catégories, par exemple, on construit un tableau de taille $n \times p$ de terme général k_{iu}, tel que :

$$k_{iu} = \begin{cases} 1 & si\ i\ appartient\ à\ la\ catégorie\ u \\ 0 & dans\ le\ cas\ contraire \end{cases}$$

Ce tableau ayant la propriété d'unicité suivante : $\sum_{u=1}^{p} k_{iu} = k_{i.} = 1 \quad \forall i$

Le nombre d'individus appartenant à la catégorie u de la variable V est donné par : $\sum_{i=1}^{n} k_{iu} = k_{.u} = n_{u.} \qquad \forall u .$

La relation scalaire qui existe entre le tableau disjonctif et le tableau de comparaisons par paires de la variable V est donnée par : $\sum_{u=1}^{p} k_{iu} k_{i'u} = c_{ii'} \quad \forall i, i'$.

Nous verrons par la suite que le passage au tableau disjonctif complet sera très pratique pour démontrer les liaisons existantes entre l'approche contingentielle et l'approche comparaisons par paires.

Grâce aux relations mathématiques, que nous venons de voir, on peut déduire que :

$$\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{u=1}^{p} k_{iu} k_{i'u} = \sum_{i=1}^{n}\sum_{i'=1}^{n} c_{ii'} = c_{..}$$

$$= \sum_{u=1}^{p}\left(\sum_{i=1}^{n} k_{iu}\right)\left(\sum_{i'=1}^{n} k_{i'u}\right) = \sum_{u=1}^{p} n_{u.}^2$$

2.8.2 Propriétés relatives à deux variables

Supposons que deux variables qualitatives V et X, ayant respectivement p et q catégories sont mesurées sur la population. Si k_{iu}^1 et k_{iv}^2 sont respectivement les termes généraux des tableaux disjonctifs complets relatifs aux codages binaires de V et X, alors le nombre d'individus n_{uv} dans la population qui appartiennent à la fois à la catégorie u de V et à la catégorie v de X est donné par : $n_{uv} = \sum_{i=1}^{n} k_{iu}^1 k_{iv}^2$.

On peut montrer que : $\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 = \sum_{i=1}^{n}\sum_{i'=1}^{n} c_{ii'} x_{ii'}$, où, comme nous le supposerons dans la suite de cet article :

- $c_{ii'}$ et $x_{ii'}$ sont respectivement les termes généraux des tableaux relationnels correspondants aux variables V et X.

en effet :

$$\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 = \sum_{u=1}^{p}\sum_{v=1}^{q}\left(\sum_{i=1}^{n} k_{iu}^1 k_{iv}^2\right)^2 = \sum_{u=1}^{p}\sum_{v=1}^{q}\left(\sum_{i=1}^{n} k_{iu}^1 k_{iv}^2\right)\left(\sum_{i'=1}^{n} k_{i'u}^1 k_{i'v}^2\right)$$

$$= \sum_{i=1}^{n}\sum_{i'=1}^{n}\left(\sum_{u=1}^{p} k_{iu}^1 k_{i'u}^1\right)\left(\sum_{v=1}^{q} k_{iv}^2 k_{i'v}^2\right) = \sum_{i=1}^{n}\sum_{i'=1}^{n} c_{ii'} x_{ii'}$$

Les propriétés que nous venons de voir, montrent clairement que des formules non linéaires dans l'espace des contingences deviennent linéaires dans l'espace des comparaisons par paires.

2.8.3 Linéarisation du critère de Belson

Sous sa forme contingentielle développée, le critère de Belson, $B(V,X)$ pour deux variables V et X est donné par la relation :

$$B(V,X)=\sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv}^2-2n_{uv}\frac{n_{u.}n_{.v}}{n}+\frac{n_{u.}^2n_{.v}^2}{n^2}\right)$$

Pour obtenir la formulation relationnelle de ce critère, nous devons tout d'abord trouver la formulation relationnelle de la quantité : $\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}n_{u.}n_{.v}$. En remplaçant les trois termes contingentiels de cette double somme par leurs expressions relationnelles définies ci-dessus, on obtient :

$$\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}n_{u.}n_{.v}=\sum_{u=1}^{p}\sum_{v=1}^{q}\left(\sum_{i=1}^{n}k_{iu}^1k_{iv}^2\right)\left(\sum_{i"=1}^{n}k_{i"u}^1\right)\left(\sum_{i'=1}^{n}k_{i'v}^2\right)$$

après permutations des sommations et changement de positions des termes on obtient :

$$\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}n_{u.}n_{.v}=\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{i"=1}^{n}\left(\sum_{u=1}^{p}k_{iu}^1k_{i"u}^1\right)\left(\sum_{v=1}^{q}k_{iv}^2k_{i'v}^2\right)$$

$$=\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{i"=1}^{n}c_{ii"}x_{ii'}=\sum_{i=1}^{n}\sum_{i'=1}^{n}c_{i.}x_{ii'}$$

Comme V est une partition elle est symétrique et donc $c_{i.}=c_{.i}$, on en déduit :

$$\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}n_{u.}n_{.v}=\sum_{i=1}^{n}\sum_{i'=1}^{n}\frac{(c_{i.}+c_{.i'})}{2}x_{ii'}$$

en remplaçant, dans le critère de Belson, tous les termes contingentiels par leur équivalents relationnels on peut facilement montrer que :

$$B(V,X)=\sum_{i=1}^{n}\sum_{i'=1}^{n}\left(c_{ii'}-\frac{c_{i.}+c_{.i'}}{n}+\frac{c_{..}}{n^2}\right)x_{ii'}$$

Sous cette forme, on voit que le coefficient de $x_{ii'}$ n'est rien d'autre que la « décomposition de Torgerson » du tableau carré relationnel correspondant à la variable V . On montre donc que le critère de Belson est bien une fonction linéaire de $x_{ii'}$.

2.8.4 Linéarisation du critère de Rand

Le critère de Rand sous sa forme développée, s'écrit :

$$R(V,X)=\frac{2\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}^2+\sum_{u=1}^{p}n_{u.}^2+\sum_{v=1}^{q}n_{.v}^2+n}{n^2}$$

en utilisant les relations précédentes, on montre que :

$$R(V,X)=\frac{\sum_{i=1}^{n}\sum_{i'=1}^{n}(2c_{ii'}x_{ii'}-c_{ii'}-x_{ii'}+1)}{n^2}$$

en posant $\overline{x}_{ii'}=1-x_{ii'}$ et $\overline{c}_{ii'}=1-c_{ii'}$, ce critère devient :

$$R(V,X)=\frac{\sum_{i=1}^{n}\sum_{i'=1}^{n}(c_{ii'}x_{ii'}+\overline{c}_{ii'}\overline{x}_{ii'})}{n^2}$$

Grâce à cette écriture, on voit que le critère de Rand (introduit en 1971), n'est rien d'autre que le critère de Condorcet (introduit en 1785) divisé par n^2.

Beaucoup d'autres critères contingentiels, ont été transformés en leurs équivalents relationnels par F. Marcotorchino (1984).

2.8.5 Intérêt majeur de la linéarisation contingentielle : l'Association maximale

En effet, si l'on prend, par exemple, le critère de Rand comme critère d'association entre variables qualitatives et que l'on cherche une partition inconnue X (variable qualitative) telle que : $R(V^1,X),R(V^2,X),\ldots,R(V^m,X)$ soient calculées simultanément, du fait de la dernière représentation relationnelle du critère de Rand, on peut chercher X telle que $\sum_{k=1}^{m}R(V^k,X)$ soit maximum. Sans la fixation a priori du nombre de classes de la partition inconnue X, ce problème est non calculable et non modélisable dans l'espace contingentiel. En revanche, en transformant, pour $(k=1,2,\ldots,m)$ les $R(V^k,X)$ par leurs formulations relationnelles, on obtient :

$$\sum_{k=1}^{m}R(V^k,X)=\frac{1}{n^2}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{i'=1}^{n}(c_{ii'}^{k}x_{ii'}+\overline{c}_{ii'}^{k}\overline{x}_{ii'})=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}(c_{ii'}x_{ii'}+\overline{c}_{ii'}\overline{x}_{ii'})$$

$$=\frac{1}{n^2}C(X)$$

On voit donc que la maximisation de $\sum_{k=1}^{m}R(V^k,X)$ revient, au coefficient $\frac{1}{n^2}$ à maximiser le critère de Condorcet qui a l'avantage majeur de ne pas avoir à fixer le nombre de classes de la partition inconnue X.

Si l'on remplace le critère de Rand par tout autre critère contingentiel, il suffit de remplacer ce dernier par une formulation relationnelle Λ de ce même critère linéaire en $x_{ii'}$ sous la forme : $Max\left(\sum_{k=1}^{m}\Lambda(V^k,X)\right)$ sous la contrainte X relation d'équivalence. Ce pro-

blème, dit « d'association maximale » est donc une variante du problème de classification relationnelle.

3 Application à des données réelles

Nous avons utilisé notre outil de classification RaresText, basé sur la théorie relationnelle, pour classifier la base de données textuelle « 20 Newsgroups » qui est devenue une référence sur laquelle des techniques différentes de fouille de données sont utilisées par la communauté scientifique et technique. Cette base est constituée de 19 997 documents, issus de 20 forums différents et décrits par 145 980 descripteurs. La caractéristique essentielle de cette base est son hétérogénéité à la fois en termes de taille des documents, de leurs thématiques ainsi que de leurs styles. Les détails du processus utilisé peuvent être consultés dans Lemoine et al. (2006).

Nous donnerons à titre d'exemple, la liste des 7 premières grandes classes obtenues en explicitant les descripteurs qui ont le plus participé à leur constitution ainsi qu'un essai d'interprétation des thématiques traitées par les documents constituants ces classes.

3.1 Echantillon du résultat obtenu

Nous avons obtenu, à l'issu du processus de classification, 330 classes. Ces classes ont été triées par ordre décroissant de leur effectif (cardinal).

Classe	Descripteurs	Cardinal
1	game, team, player, hockey, season, playoff, fan, baseball, league, coach	1325
2	file, directory, program, window, FTP, archive, DOS, disk, server	1144
3	Government, right, law, constitution, weapon, citizen, president, gun, policy	1095
4	Car, engine, mile, tire, mileage, brake, dealer, wheel, auto, clutch	755
5	Clipper, encryption, key, chip, escrow, crypto, wire tap, algorithm, privacy, government	673
6	Drive, SCSI, IDE, disk, controller, ram, floppy, CD-ROM, jumper, software	628
7	Card, video, driver, ISA, monitor, bus, VGA, VLB, SVGA, graphics	579

TAB. 6 – Les sept premières classes de la partition finale.

3.2 Essai d'interprétation

On peut observer, au vu des descripteurs caractérisant les classes :, que la classe 1 traite du « sport » en général ; la classe 2 du « logiciel » ; la classe 3 de la « politique » ; la classe 4 de « l 'automobile » ; la classe 5 du « cryptage » et de la protection des données ; la classe 6

du « matériel informatique » en général et plus particulièrement du choix IDE ou SCSI et enfin la classe 7 traite aussi du « matériel informatique » en général et plus particulièrement du matériel vidéo.

3.3 Conclusion

La théorie relationnelle nous a permis d'obtenir une classification du corpus « 20 NewsGroup » sans avoir recours à de l'échantillonnage ni à la fixation a priori du nombre de classes pouvant exister dans le corpus. Notre technique, au vu des résultats obtenus, nous a permis de mettre en évidence, à la fois, les grandes thématiques générales du corpus et des sous-thématiques plus spécifiques. Nous avons aussi découvert des classes de documents issus de plusieurs forums, identifiant ainsi des liens cachés entre ces derniers.

Références

Benhadda, H. (1998) La similarité régularisée et ses applications en classification automatique. Thèse de doctorat, Paris VI.

Benhadda, H., F. Marcotorchino (1997). Introduction à la similarité régularisée an analyse relationnelle (cas qualitatif). *Congrès des 29ème Journées de Statistique (Carcassonne 26-30 Mai),* 136–138.

Benhadda, H., F. Marcotorchino (1998). Introduction à la similarité régularisée an analyse relationnelle. *Revue de statistique appliquée Vol. 46 N°1,* 45–69.

Kendall, M. G., B. Babington Smith (1940). On the method of paired comparisons. *Biometrika 31*.

Lemoine, J., H. Benhadda et J. Ah-Pine (2006). Classification non supervisée de documents hétérogènes : Application au corpus « 20 NewsGroups ». 11th *IPMU*.

Marcotorchino, F. (1984) Utilisation des comparaisons par paires en statistique des contingences: partie I. *Etude du centre scientifique IBM France F-069*.

Marcotorchino, F. (1984) Utilisation des comparaisons par paires en statistique des contingences: partie II. *Etude du centre scientifique IBM France F-071*.

Marcotorchino F. (1987). Block seriation problems: A unified approach. *Applied stochastic models and data analysis* 3, 73–91.

Marcotorchino F. (1991). Seriation problems: An overview. *Applied stochastic models and data analysis* 7, 139–151.

Marcotorchino, F. (1991) L'analyse factorielle-relationnelle : parties I et II. *Etude du centre scientifique IBM France MAP-03*.

Marcotorchino, F., P. Michaud (1978). *Optimisation en analyse ordinale des données*. Masson.

Michaud, P. (1981) Agrégation de préférences. Thèse de doctorat, Paris VI.

Michaud, P. (1985) Agrégation à la majorité II : analyse du résultat d'un vote. *Etude du centre scientifique IBM France F.094*.

Rand W. H. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Associations* 66.

Summary

In this article we will briefly show the possibilities offered by the Relational Analysis Theory (initiated in early 1980's at IBM Corp). Presently, we will give a short overview on theoretical and methodological advances obtained with this approach to merge information, to treat and analyse huge amount of data (either unstructured or structured data). We will show as well, associated transfer formulas allowing to express well known combinatorial problems into linear economical functions suitable for different kinds of problematic (such as Clustering problems, assignment problems, bi-dimensional classification, etc.). This, in addition to the O (N) order of magnitude for the computational algorithmic part, allows this approach to be tractable and pertinent for various real life applications.