Robustesse des partitions de textes : une exploration autour de l'apport des motifs de mots

Martine Cadot [1], Michel Zitt [2], Gabriel Meurin [1], Alain Lelu [1, 3]

[1] LORIA – Campus scientifique – BP 239 – 54506 Vandoeuvre cedex – France

[2] INRA SAE2 LERECO – Rue de la Géraudiere – BP 71627 – 44316 Nantes cedex - France

[3] LASELDI, 30 rue Mégevand – 25030 Besançon cedex

Résumé

Les partitions d'ensembles de textes peuvent être d'origine endogène – provenir d'une classification non supervisée – ou exogène, par catégorisation a priori des textes par des experts. Dans les deux cas se pose la question du caractère généralisable des catégories qu'elles expriment. Dans le cadre de la recherche d'une meilleure catégorisation des domaines scientifiques, et à partir d'un extrait significatif de la base de référence Web of Science, nous avons appliqué la méthodologie classique de l'apprentissage automatique (sous-corpus distincts: apprentissage, ajustement, test) à une partition non supervisée du domaine de la génomique. Les résultats sur cet ensemble de 120.000 résumés d'articles font la preuve d'une qualité et d'une robustesse accrues quand on caractérise chaque partition, plutôt que par de simples termes, par des motifs de termes. Ces motifs spécifiques de chaque catégorie sont extraits par notre méthode Midova de sélection et détermination de liaisons complexes entre variables booléennes par « pulvérisation » des effectifs impliqués dans chaque relation n-aire.

Mots-clés : analyse de données textuelles, catégorisation de textes, validation, généralisation, robustesse de partition, classification, stabilité de classification, motifs de mots, nomenclature, requête Booléenne, expansion de requête

Abstract

Text categorization may follow from an endogenous origin, in the case of unsupervised classification, or from an exogenous one when experts have tagged a sample of the corpus. In each case one may investigate the generalization potential of such categories. Trying to enhance a present categorization of scientific domains, we have indexed a significant sample of the Web of Sciences dealing with genomics, extracting lemmas and multiterms out of the abstracts. We first clusterized it into 50 categories, and we eventually submitted the resulting classes to the classic machine learning methodology, splitting the database into three sub-corpuses for training, adjusting the parameters, and validating. Our results on this 120.000 document database show that the best generalizing ability proceeds from using a compact and limited set of term itemsets rather than many simple terms, provided that those itemsets follow from our MIDOVA method for selecting and bringing to light the right complex Boolean combinations of initial binary variables.

1. Introduction et problématique

Les études bibliométriques et scientométriques utilisent les « réservoirs » de millions de références bibliographiques que sont les bases de données scientifiques, comme le Web of Science (http://thomsonreuters.com/products_services/science), Scopus (http://info.scopus.com), etc. pour observer et évaluer les mouvements des thématiques scientifiques et techniques; des

indicateurs (http://www.obs-ost.fr/fr/l-offre.html) par nations, par domaine, par institution, et de plus en plus par laboratoire, voire par chercheur, en sont tirés, avec de forts enjeux collectifs et individuels. C'est dire à quel point la question du *contour* des catégories utilisées par ces indicateurs et ces études est stratégique, et fait l'objet d'une attention de plus en plus vigilante de la part de tous les acteurs impliqués, les contestations et polémiques portant souvent sur les limites des bases de données choisies et des domaines qu'elles définissent. On distingue souvent deux types de catégorisation, mais qui n'ont rien d'absolu :

- *les nomenclatures*, outils « macro » qui supposent un choix de stabilité temporelle à moyen terme, au moins à court terme. Les documents sont assignés aux catégories, soit manuellement, avec une assistance ordinateur possible, par des experts qui indexent chaque article de revue : c'est le modèle Pascal-INIST (http://www.inist.fr), par exemple ; soit par assignation en bloc, journal par journal, à des catégories définies comme des ensembles de journaux ou sections de journaux – c'est le modèle Thomson-Reuters pour le Web of Science. Dans ce dernier cas, la supervision est limitée à la refonte annuelle de la nomenclature de journaux.
- *les classifications*, outils « micro » qui visent en général une granulométrie plus fine – les thématiques ou fronts de recherche – et une imagerie dynamique des réseaux scientifiques. La stabilité des classes n'est pas un donné mais une caractéristique locale, vérifiée ou non. Les classifications sur les graphes de liens bibliométriques (relations lexicales, citations, copublications, etc.) sont automatisables et souvent non supervisées, en dehors des phases de validation. Notre problématique se situe dans le cadre de l'extraction automatique ou semi-automatique (Bassecoulard et al., 2007) de ces catégories : quelle validité, stabilité, pérennité attribuer à ces catégories ? Comment définir de la façon la plus condensée possible ces catégories et les attribuer à tout nouvel article, par exemple en se basant sur les mots du titre et du résumé, ou sur les citations reçues ou émises ?

Notre démarche en ce sens a été résolument exploratoire : nous sommes partis d'un corpus réel de 120.000 résumés d'article du Web of Science portant sur le domaine de la génomique, dont les contours ont été obtenus par la méthode hybride « Citlex » (lexicale/citations) décrite dans (Zitt et al., 2006), l'application génomique s'étant opérée en interaction avec un panel d'experts du domaine (Laurens et al., 2008). Ce corpus a été segmenté en 50 sous-domaines de la génomique par la méthode de classification non-supervisée K-means axiales (Lelu, 1994), dont une des spécificités est d'accompagner l'attribution d'un article (décrit en l'occurrence par son résumé) à une classe par une valeur d'un indicateur de centralité de l'article dans cette classe.

Alors que la stabilité des résultats d'analyses factorielles a été étudiée par (Lebart, 2007) en exploitant statistiquement des données bruitées dérivées par bootstrap des données d'origine, nous avons choisi ici d'étudier la stabilité et le caractère général de notre partition selon la méthodologie classique utilisée en apprentissage automatique : 1) déterminer tout d'abord la fonction d'attribution des classes aux documents sur un sous-ensemble dit d'apprentissage ; 2) affiner ses paramètres sur un sous-ensemble dit de validation ; 3) enfin l'éprouver sur un 3^{e} ensemble dit de test. Nous présentons (sections 4 et 5) une combinaison originale de nos travaux antérieurs, en premier lieu en matière de création et sélection de nouveaux attributs créés par combinaisons Booléennes des attributs d'origine : chaque classe se trouve décrite par une liste ordonnée et évaluée de descripteurs anciens et nouveaux. On constate que pour l'ensemble des classes la taille totale des listes (obtenues à l'optimum du critère de qualité en généralisation) est d'un ou deux ordres de grandeur au dessous du nombre initial d'attributs. En second lieu (section 5) nous avons comparé, à nombre de descripteurs égal, nos résultats (obtenus par notre méthode au moyen de combinaisons non-linéaires d'attributs), à ceux obtenus avec des

combinaisons linéaires de simples attributs d'origine (K-means Axiales) : il ressort clairement que compacité et performance en généralisation vont de pair.

2. Les données

Elles ont été recueillies dans le cadre du projet CSTG, soutenu par l'Agence Nationale de la Recherche, en 2007. Il s'agit d'un extrait de la base Web of Science, constitué d'« articles » et « lettres » publiés de 1999 à 2005. Un traitement en plusieurs étapes a été mis en œuvre :

- Une définition de la génomique a été établie par un groupe d'experts, ainsi qu'une liste des 26 revues « cœur » du domaine, transversale à plusieurs catégories Thomson, qui a permis d'extraire 15.900 documents. Aux yeux de l'ensemble des experts consultés, ce qui distingue la génomique, au sein de la génétique, est l'échelle de l'approche. La génomique est une approche à grande échelle et ne se focalise pas, contrairement à d'autres domaines de la génétique, au gène ou au petit groupe de gènes.
- Une requête lexicale a été établie, en premier lieu à partir de termes standards de la génétique (*dna*, *rna*, ...) spécifiés par des termes caractéristiques des études à grande échelle (*large_scale*, *high_throughput*, …), en second lieu à partir de termes spécifiques de la génomique (*genebank*, *metabolom**, *bioinformatic**, ...). Ce qui a ajouté 36 600 documents au corpus.
- Un processus d'extension à partir des citations, selon la méthode définie dans (Zitt et al., 2006), a suggéré une nouvelle liste de documents, structurée selon deux indices de pertinence bibliographique, puis filtrée par les experts, ajoutant 67.200 documents au corpus précédent, dont la version définitive a atteint dès lors 119.700 documents.

Les titres et résumés de ce corpus ont été chargés dans l'environnement d'extraction de termes simples et composés et de catégorisation automatique NeuroNav (www.diatopie.com). Comparée à une extraction manuelle de termes pertinents, celle fournie par NeuroNav offre un rappel de l'ordre de 50%, suffisant compte tenu de la forte redondance présente et une précision supérieure à 95% grâce à l'élimination des mots et termes composés à caractère syntaxique ou rhétorique (*encouraging_results*, ...). Seuls les substantifs et groupes nominaux ont été retenus. Après unification complémentaire de variantes de termes, puis suppression des termes de fréquences 1 et 2, ainsi que de quatre unitermes de très haute fréquence, la taille du vocabulaire s'est établie à 203.000 termes.

3. Centralités des lignes et colonnes d'une matrice binaire et de ses parties

Notations : majuscules et minuscules grasses pour les matrices et les vecteurs ; **X'** est la transposée de **X**, matrice de I lignes et T colonnes. Les sommes en lignes et en colonnes de ses éléments x_{it} s'écrivent respectivement $x_{i.}$ et $x_{.t}$, sa somme totale $x_{..}$.

3.1. Distance et cosinus distributionnels

Une lignée ancienne de travaux (Matusita, 1955 ; Escofier, 1978 ; Domengès and Volle, 1979 ; Rao, 1995) s'est intéressée à ce que certains auteurs appellent distance distributionnelle et d'autres distance de Hellinger : il s'agit de la distance euclidienne, classique (équipondération des dimensions), entre les deux points t et t', de coordonnées les vecteurs $\mathbf{z}_t$ et $\mathbf{z}_{t'}$, situés sur l'hypersphère unité dans l'espace des I mots et représentant chacun une unité de découpage textuel, définis par la transformation suivante sur les données : $\mathbf{z}_t$: $\{\sqrt{(x_{it}/ x_{.t})}\}$ où x_{it} désigne la fréquence du mot i dans le document t, et $x_{.t}$ le nombre total de mots du document t.

La distance distributionnelle Dd(t, t') entre les textes t et t' est donc :

$$Dd(t, t') = \| \mathbf{z}_t - \mathbf{z}_{t'} \| \text{ où } \| \mathbf{.} \| \text{ désigne la norme euclidienne d'un vecteur.}$$

Cette distance est la longueur de la corde correspondant à l'angle $(\mathbf{z}_t, \mathbf{z}_{t'})$ – égale au plus à 2 quand ces deux vecteurs normalisés sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux. Cette distance semble triviale et arbitraire en apparence (pourquoi cette normalisation insolite plutôt que la normalisation classique $\{x_{it} / \|\mathbf{x}_t\|\}$?), mais elle jouit de propriétés intéressantes :

- Contrairement à la distance du khi-deux utilisée en Analyse Factorielle des Correspondances (AFC), elle peut prendre en compte des vecteurs ayant des composantes négatives, propriété utile pour certains types de codage « symétriques » (comme Oui, Non, Ne sait pas) ou pour des tableaux de flux orientés – économiques, physiques, ... ;
- Elle est liée à la mesure du gain d'information de Renyi (1966) d'ordre ½ apporté par une distribution $\mathbf{x}_q$ quand on connaît la distribution $\mathbf{x}_p$:
$$I^{(1/2)}(\mathbf{x}_q / \mathbf{x}_p) = - 2 \log_2 (\cos(\mathbf{z}_p, \mathbf{z}_q)) = - 2 \log_2 (1 – Dd^2/2) ;$$
- Elle est particulièrement adaptée aux « données directionnelles » (Banerjee et al., 2005) que sont les données textuelles, pour lesquelles seuls sont pertinents les angles entre vecteurs ;
- Elle est rapide à calculer dans le cas des données textuelles, où les vecteurs $\mathbf{x}_t$ sont creux ;
- et surtout Escofier et Volle ont montré qu'elle satisfaisait à la même propriété d'équivalence distributionnelle que la distance du khi-deux utilisée en AFC: si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, dans le cas où les descripteurs sont des mots et les unités décrites des textes, cette propriété assure la stabilité du système des distances entre textes au regard de l'éclatement ou du regroupement de mots de distributions proches.

3.2. Indicateurs de centralité spectrale pour les lignes et les colonnes

L'analyse factorielle sphérique (AFS) (Domengès and Volle, 1979), dans son option "différence par rapport au tableau nul", consiste à réaliser la décomposition aux valeurs singulières du nuage de points $\{ \mathbf{z}_t \}$ pondérés par les sommes marginales $x_{.t}$; ou de façon duale (on montre que leurs valeurs singulières sont les mêmes) celle du nuage de points $\{ \mathbf{z}_i \}$ pondérés par les sommes $x_{i.}$. Elle se passe donc dans les deux espaces de Hellinger duaux : l'hypersphère unité des lignes et celle des colonnes.

Soit $\mathbf{X}^{1/2} = \mathbf{U\ D\ V'}$ la décomposition aux valeurs singulières de la matrice $\mathbf{X}^{1/2}$ des racines carrées des valeurs de $\mathbf{X}$ (ce sont les mêmes valeurs quand les données sont binaires). Les facteurs sphériques $\mathbf{F}$ et $\mathbf{G}$ (orthonormés) s'en déduisent : $\mathbf{F} = \mathbf{Dr}^{-1/2}\ \mathbf{U\ D}$; $\mathbf{G} = \mathbf{Dc}^{-1/2}\ \mathbf{V\ D}$ où $\mathbf{Dr}$ et $\mathbf{Dc}$ sont les matrices diagonales des sommes en lignes et colonnes de $\mathbf{X}$. D'où la formule de reconstitution des données en AFS : $\mathbf{X}^{1/2} = \mathbf{Dr}^{1/2}\ \mathbf{F\ D}^{-1}\ \mathbf{G'\ Dc}^{1/2}$

Et les formules de transition entre facteurs AFS :

$$\mathbf{F} = \mathbf{Dr}^{-1/2}\ \mathbf{X}^{1/2}\ \mathbf{Dc}^{1/2}\ \mathbf{G\ D}^{-1} \ ; \qquad \mathbf{G} = \mathbf{Dc}^{-1/2}\ (\mathbf{X}^{1/2})'\ \mathbf{Dr}^{1/2}\ \mathbf{F\ D}^{-1}$$

Cette analyse est formellement proche de l'AFC, où une SVD est réalisée sur la transformée $\mathbf{Q}$ de la matrice des données : $\mathbf{Q} = \mathbf{Dr}^{-1/2}\ \mathbf{X}\ \mathbf{Dc}^{-1/2}$

Ce qui mène à la décomposition : $\mathbf{Q} = \mathbf{U}_{afc}\ \mathbf{D}_{afc}\ \mathbf{V}_{afc}'$

Et aux facteurs AFC : $\mathbf{F}_{afc} = x_{..}^{1/2}\ \mathbf{Dr}^{-1/2}\ \mathbf{U}_{afc}\ \mathbf{D}_{afc}$; $\mathbf{G}_{afc} = x_{..}^{1/2}\ \mathbf{Dc}^{-1/2}\ \mathbf{V}_{afc}\ \mathbf{D}_{afc}$

Et à la reconstitution des données : $\mathbf{X} = x_{..}^{-1}\ \mathbf{Dr}\ \mathbf{F}_{afc}\ \mathbf{D}_{afc}^{-1}\ \mathbf{G}_{afc}'\ \mathbf{Dc}$.

Alors que le premier facteur AFC est trivial (vecteur unitaire), il découle immédiatement de l'identité des vecteurs singuliers entre toute matrice $\mathbf{Z}$ et la matrice $\mathbf{Z'\ Z}$ qui en dérive que le 1er

facteur AFS est le même que celui de la matrice carrée des cosinus entre lignes $\mathbf{X}^{1/2}\ \mathbf{Dc}^{1}\ (\mathbf{X}^{1/2})'$ qu'on peut interpréter comme la matrice d'adjacence du graphe évalué que forment les lignes entre elles. Les valeurs des lignes pour ce premier facteur peuvent donc être interprétées comme les centralités « spectrales » (Brandes and Cornelsen, 2003) des nœuds de ce graphe ; et de façon duale pour les centralités des colonnes. Ainsi des vecteurs-lignes angulairement proches de nombreux autres, centraux dans le graphe, auront une valeur forte de cet indicateur, indépendamment des valeurs de leurs composantes dans l'absolu, et les vecteurs-lignes isolés, périphériques, auront une valeur faible.

3.3. Extension aux catégories d'une partition

Le calcul des centralités décrit ci-dessus peut être appliqué à chacune des K partitions d'un ensemble de vecteurs, représentant par exemple des sous-ensembles de textes homogènes, comme dans notre exemple d'application. Dans ce dernier cas, la méthode de clustering qui a été choisie, les K-means axiales, utilise précisément les indicateurs de centralité décrits ci-dessus comme critères d'attribution d'un individu (ici, un résumé d'article) à un cluster ; géométriquement, cet indicateur représente le cosinus entre le vecteur-document et le vecteur-axe de cluster, et comme ces vecteurs sont unitaires, il représente la projection du premier sur le deuxième. Le maximum des projections sur les 50 axes de cluster détermine le cluster d'appartenance. Nous exposerons en section 5 comment nous avons testé le caractère généralisable de nos 50 segments de matrice à partir de tout ou partie de ces descripteurs.

Pour tester ce caractère dans le cas où la partition est établie sur des critères exogènes au corpus, nous avons recueilli la totalité des 68.000 résumés d'un mois de WoS (toutes disciplines) dans la période 1999-2005, et marqués les 916 appartenant au domaine génomique inclus dans notre premier corpus, ensemble défini par une procédure complexe supervisée par des experts. Mais la disproportion entre le nombre des documents des deux classes, et la taille insuffisante de l'extrait « genomique » de l'échantillon au regard de son nombre d'attributs, nous a empêché d'appliquer dans un délai raisonnable la méthodologie classique de l'apprentissage. Ce travail fait partie de nos perspectives.

4. Décomposition MIDOVA d'une matrice binaire

4.1. Contexte et motivation

Nous nous plaçons dans le cas d'une relation R entre deux ensembles, un ensemble S d'individus (ou sujets, instances, transactions) s_i, et un ensemble V de variables binaires (ou attributs, propriétés) v_j. Dans notre cas, les individus sont des textes, et les variables sont les termes qu'ils contiennent ou non. Selon les diverses sources de données, la relation binaire R est fournie sous forme de listes – liste des termes des textes – ou sous forme d'un tableau ayant pour lignes les individus, pour colonnes les variables, avec la valeur 1 à l'intersection de la ligne i et de la colonne j si l'individu s_i possède la propriété v_j – le texte i contient le terme j – et la valeur 0 dans le cas contraire. Avec MIDOVA, nous proposons une autre écriture de la relation R, sous forme d'une liste de relations entre variables de l'ensemble V. Dans cette écriture, l'ensemble S a disparu, les individus étant remplacés par les liens qu'ils ont tissés entre les variables à travers la relation R. Plus les liens tissés sont nombreux et forts, moins il y a de relations dans la liste, et moins elles impliquent de variables. Le nombre de composants de la liste varie ainsi entre p et 2^p, où p est le nombre de variables de V. Le premier cas correspond à la liaison la plus forte qui puisse exister : tous les individus sont semblables, du moins pour ces variables qui deviennent des constantes (0 pour tous les individus, ou 1 pour tous), et la connaissance

de la valeur de chaque variable suffit à définir R. Le second cas correspond à la liaison la plus faible qui puisse exister entre les variables : la connaissance des valeurs d'un individu pour k variables ne permet pas de deviner la valeur pour une k+1ème variable, et ceci pour tout k. Le nombre maximum 2^p de relations ne peut toutefois être atteint que si le nombre d'individus N est supérieur à 2^p. Entre ces deux extrêmes, le nombre d'éléments de la décomposition de R peut varier énormément, mais généralement le nombre de variables de chacun de ses éléments reste raisonnable. Un paramètre de MIDOVA permet de choisir le niveau d'approximation que l'on souhaite. Si on n'autorise aucune approximation la reconstitution de la relation R de départ est exacte à une renumérotation des individus près. Plus la tolérance est grande, plus la liste est petite, et moins il y a de variables dans ses éléments, mais plus la reconstitution de la relation R de départ est approximative.

4.2. Principes et méthode

La représentation MIDOVA (Cadot, 2006) s'appuie sur le concept statistique de "table de contingence" entre deux variables a et b (pour une vue d'ensemble, cf. Morineau, 1996), généralisée en une "hyper-table de contingence" pour plus de deux variables. La table de contingence classique est formée de quatre cellules (pour les variables booléennes) dans lesquelles se répartissent la totalité des N individus selon leurs valeurs dans la paire de variables {a,b}. Les quatre valeurs possibles de la paire sont a=1 et b=1, a=1 et b=0, a=0 et b=1, et a=0 et b=0, notées plus rapidement ab, $a\bar{b}$, $\bar{a}b$, $\bar{a}\bar{b}$. On peut voir dans la figure 2 un exemple de table de contingence montant la répartition de 100 individus en 20, 6, 14 et 60 selon les 4 valeurs de la paire {a,b}. Pour 3 variables, il y a deux fois plus de valeurs, et la table de contingence contient 2^3 cellules disposées dans un cube. Chaque nouvelle variable double le nombre de cellules de la table de contingence. Comme la somme de toutes les cellules d'une table de contingence est N, le contenu des cellules est en moyenne de plus en plus petit, et il apparaît de plus en plus de cellules vides. Notons que nous pouvons étendre également la notion de tableau de contingence de dimension k à des valeurs de k inférieures à 2 (Fig. 1), une table de dimension 1 ayant deux cellules et correspondant à une seule variable, comme "a" (effectifs de a et $\bar{a}$) ou "b", la table de dimension 0 ayant une seule cellule qui contient le nombre total d'individus et correspond à un ensemble vide de variables.

La propriété sur laquelle s'appuie la décomposition MIDOVA est la suivante : une table de contingence de dimension k exprimant une relation entre k variables n'a qu'un « seul degré de liberté » si on connaît toutes les tables de contingence de dimension k-1 correspondant à toutes les combinaisons de k-1 variables parmi les k variables. De plus c'est une « liberté sous contraintes ». Cette propriété est illustrée dans les figg. 1 et 2.

n
100

	a	$\bar{a}$	Tot.
	26	74	100

	b	$\bar{b}$	Tot.
	34	66	100

	b	$\bar{b}$	Tot.
a	26-x	x	26
$\bar{a}$	74-(66-x)	66-x	74
Tot.	34	66	100

Figure 1 : Création d'une table de contingence de dimension 2 à partir de celles de dimensions inférieures

Dans Fig. 1, on peut voir qu'une fois connu le contenu "x" d'une cellule de la table de contingence de (a,b), le contenu des trois autres cellules s'obtient à partir de cette valeur "x" et des valeurs des tables de contingence de (a) et de (b), et "x" ne peut pas prendre de valeurs en dehors de l'intervalle [0,26]. Dans Fig. 2, on fait de même avec la table de (a, b, c) connaissant celles de (a, b), (a, c), (b, c), et on remarque que x est astreint à prendre ses valeurs dans l'intervalle [0,6].

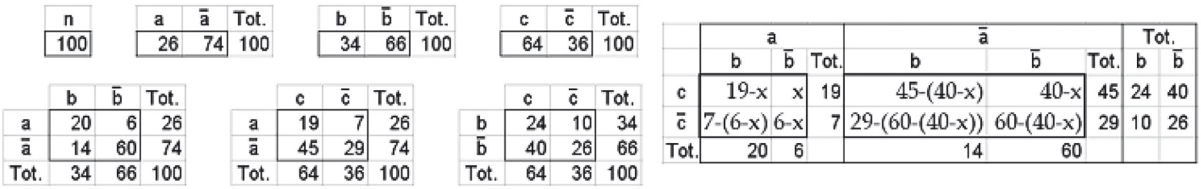

Figure 2 : Création d'une table de contingence de dimension 3 à partir de celles de dimensions inférieures

Il découle de cette propriété les conséquences suivantes : en premier lieu la connaissance d'une seule cellule des 2^p tables de contingence (une par sous-ensemble de variables de V) suffit à décrire la relation R. On choisira pour chaque ensemble de variables l'effectif correspondant à leurs valeurs toutes égales à 1. Par exemple, si x=0 dans Fig. 2, la relation entre a, b et c est entièrement décrite par Ø (100), a(26), b(34), c(64), ab(20), ac(19), bc(24), abc(19). Ensuite, dès qu'une table de contingence liant k variables contient une cellule vide, toute variable ajoutée à ces k variables débouche sur une table qui n'a plus aucun degré de liberté : l'effectif total N se distribue entre les 2^k cellules de la table d'une seule façon possible qui peut se déduire des tables de dimensions inférieures par un raisonnement algébrique. Il est alors inutile de faire figurer dans la décomposition de la relation R d'autres composants contenant ces k variables. Par exemple, si R contient les 3 variables a, b et c de la figure 2 avec x=0, une des 8 cellules de la table (a, b, c) étant nulle, la relation (a, b, c) figurera dans la décomposition MIDOVA de R, mais aucune relation de plus de trois variables contenant a, b et c ne figurera dedans. Enfin quand l'effectif N est distribué dans les 2^k cellules d'une table de contingence de dimension k, si $N<2^k$ alors une cellule au moins est vide, et aucune table de dimension supérieure à k ne sera créée. Aussi le nombre maximal de variables dans un composant de R est égal à $\log_2(N)+1$.

MIDOVA est un algorithme par niveau, dérivé de l'algorithme « A priori » d'Agrawal (1999), pour lequel les sous-ensembles de variables de V sont appelés « motifs », et le « support » d'un motif est le contenu de la cellule de la table de contingence correspondant à toutes les variables simultanément vraies, cellule que nous avons choisie également dans la représentation MIDOVA de R. Le paramètre essentiel de notre algorithme est le « reste », qui évalue pour chaque motif sa capacité à créer des sur-motifs « inattendus » avec d'autres variables. En fixant le seuil de reste à 1, on obtient une décomposition exacte de R telle que décrite précédemment. Cette décomposition est unique et permet de reconstituer R aux numéros d'individus près. On peut aussi imposer un seuil de reste plus élevé, pour obtenir une décomposition de R plus approximative, mais moins sensible aux valeurs aberrantes, extrêmes, qu'elles soient dues aux erreurs de saisie, de mesure ou au hasard. Un seuil de support supérieur à 0 peut également être choisi, pour les mêmes raisons. Dans la partie suivante, nous appliquons cette méthode sur un dixième de nos données, soit (12.019 résumés d'articles en génomique tirés au hasard), contenant 95.722 termes différents.

4.3. Applications de MIDOVA aux données

Nous appliquons l'algorithme décrit dans (Cadot, 2006) avec un seuil de reste et de support de 10. Les résultats se trouvent dans la partie gauche du Tab. 1 (5 premières colonnes). À l'étape 1, on crée les 1-motifs, qui sont tout simplement les termes. Il y en a 6.112 de support supérieur ou égal à 10, c'est-à-dire présents dans au moins 10 textes. Puis on crée les 2-motifs. On en trouve 95.936 de support >=10, comportant 4.736 termes différents. Parmi ceux-ci 92.827, comportant 3.749 mots, sont de reste >=10, et vont permettre de construire les 3-motifs, et 94.000 restent pour interprétation après nettoyage (on a enlevé les motifs comportant un terme composé et l'un de ses composants).

	Nombre de termes dans les k-motifs		*Nombre de k-motifs générés (supp et reste >=10)*			*Nombre de k-motifs significatifs d'un nombre « nbcl » de clusters*			*Nb total de motifs significatifs pour les 50 clusters*
	Étape k	*Étape k+1*	*Étape k*	*Étape k+1*	*Nettoyés*	*nbcl=0*	*nbcl=1*	*nbcl>1*	
1	6.112	6.112	6.112	6.112	6.112	1.344	2.118	2.650	10.599
2	4.736	3.749	95.936	92.827	94.000	18.825	40.398	34.777	130.257
3	2.216	1.994	228.745	216.857	228.205	33.975	109.725	84.505	307.190
4	1.063	953	174.317	152.145	174.041	17.241	90.761	66.039	237.209
5	468	450	52.114	44.849	52.059	2.655	28.660	20.744	73.427
6	195	133	6.804	2.753	6.798	71	3933	2.794	9.913
7	40	21	118	10	118	0	83	35	160
8			0	0	0				
Total			*557.342*	*515.553*	*561.333*	*74.111*	*275.678*	*211.544*	*768.755*

Tableau 1 : A gauche les k-motifs créés par Midova, à droite ceux qui sont jugés significatifs pour des clusters

On a ainsi extrait d'un dixième des données une description approximative de la relation entre les textes et les termes. Elle est formée de 561.333 motifs d'au plus 7 termes chacun. Nous cherchons maintenant parmi ces k-motifs ceux qui sont significatifs de chaque cluster. Pour cela nous utilisons un test de randomisation (Manly, 1997). Chaque motif est distribué aléatoirement dans le nombre de textes (du dixième des données) dans lequel il figure, et son support est ainsi réparti entre les divers clusters. Cette opération est recommencée cent fois de façon indépendante et permet de construire un intervalle à 99% de confiance du support attendu de ce motif pour chaque cluster en cas d'affectation au hasard de ce motif aux clusters. Si le support du motif pour un cluster est supérieur à la borne supérieure de l'intervalle, il est alors jugé significatif du cluster. Dans Fig. 3, on a représenté pour chaque valeur de k la proportion de k-motifs significatifs de 0, 1, …, 15 clusters. On voit que plus k augmente, plus la proportion de k-motifs n'étant significatifs que d'un seul cluster augmente (hachures obliques), alors que diminue celle des k-motifs significatifs d'aucun cluster. Au vu de ces premiers résultats, la capacité de classement des motifs semble plus élevée pour les motifs les plus longs. Chaque k-motif extrait est en moyenne significatif de 1 à 2 clusters. On a indiqué dans la partie droite du tableau combien les motifs de chaque niveau génèrent de motifs significatifs d'aucun cluster, d'un seul cluster, ou de plus d'un cluster, et en dernière colonne le nombre de motifs significatifs de clusters qui en résultent. Le total est de 768.755 motifs significatifs comportant de 1 à 7 termes.

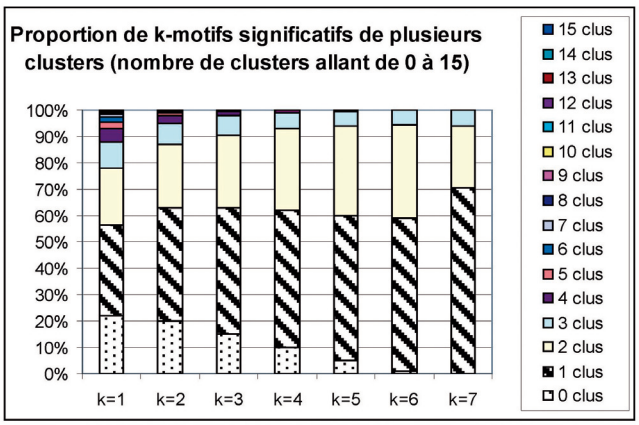

Figure 3 : Répartition des k-motifs selon le nombre de clusters dont ils sont significatifs

5. Comparaison des apprentissages à partir des p-motifs significatifs, vs. attributs d'origine centraux

La phase d'apprentissage décrite précédemment a fourni une liste de motifs significatifs des clusters à partir d'un ensemble d'entraînement constitué d'un dixième seulement des textes. Nous sommes passés ensuite à une phase d'évaluation, visant à mesurer le pouvoir de discrimination de cette liste de motifs sur le reste du corpus. Nous avons procédé en plusieurs étapes : tri de la liste des motifs, puis pour chaque texte, recherche des motifs pris selon l'ordre de la liste, calcul du cluster le plus probable selon les motifs trouvés, confrontation avec le cluster véritable du texte, évaluation de la qualité globale de prédiction pour chaque cluster et chaque dixième de corpus. Chaque étape dépend de plusieurs paramètres qui ont été ajustés lors d'une phase d'ajustement en confrontant les résultats des huit dixièmes du corpus, l'évaluation finale étant donnée sur le dernier dixième. Puis nous avons comparé les résultats de cet apprentissage à ceux obtenus à partir des attributs d'origine centraux.

5.1. Tri de la liste des motifs

Pour chaque motif M significatif d'un cluster C on a réparti les documents de l'ensemble d'entraînement en 4 parties : (a) nombre de documents de C contenant M, (b) de C ne contenant pas M, (c) en dehors de C contenant M, (d) en dehors de C ne contenant pas M. Avec ces quatre nombres, nous avons calculé la force du lien entre M et C par un indice de spécificité. Parmi la trentaine d'indices testés, l'indice qui s'est avéré le plus performant en terme de qualité de généralisation dans la phase d'ajustement est l'indice d'Ochiaï = $a/\sqrt{((a+b)(a+c))}$, corrigé de l'éventuelle nullité du dénominateur. La liste de motifs a alors été triée selon les valeurs décroissantes de cet indice.

5.2. Estimation du cluster d'appartenance d'un document

Pour ce document, le degré d'appartenance estimé pour chaque cluster est initialisé à 0. On a pris les X premiers motifs de la liste, et on a pris en compte chaque motif trouvé dans le document par une mise à jour du degré d'appartenance estimé pour le cluster correspondant. Celle-ci comportait deux étapes : calcul d'une valeur mesurant l'adéquation relative entre le motif et le cluster, prise en compte de cette valeur pour modifier le degré d'appartenance. Le cluster estimé du motif est celui qui a la plus grande valeur. Parmi les nombreuses procédures de modification du degré d'appartenance testées (valeurs binaires à additionner, maximum des rangs, etc.), on a gardé celle qui donnait le meilleur résultat dans la phase d'ajustement : mise à jour par ajout de la valeur utilisée pour classer les motifs de la liste.

5.3. Évaluation de la qualité des résultats

Pour mesurer la qualité des résultats, on a calculé le F-score (moyenne harmonique du rappel et de la précision) pour chaque cluster C à partir du décompte des documents du dernier dixième du corpus (l'ensemble de test) en « Vrais Positifs », « Vrais Négatifs », etc. Les résultats sont en figure 4. Ils montrent la grande variabilité de la qualité des clusters : certains, comme le N°45 (*LOD/ Linkage_analysis/ Polymorphism*) ou le 16 (*Spectrometry/ Proteomics*) dépassent les 80% ; d'autres, comme le 14 (*Locus/ Microsatellite_Locus/ Polymorphism*) ou le 38 (*genome/ genome_size*) n'atteignent pas 50%.

A titre d'illustration, voici un aperçu des six premiers k-motifs du cluster 10 (sur les 34 aboutissant à un AUC de 0,957) : *comparative_genomic_hybridization*, *hybridization* ET *tumor*

ET *genomic*, *hybridization* ET *tumor*, *CGH* ET *comparative_genomic_hybridization*, *losses* ET *genomic*, *tumor* ET *genomic*. Constat : le développé de l'acronyme CGH figure avec *CGH*, et ses composants sont omniprésents, seuls ou en conjonction avec d'autres mots.

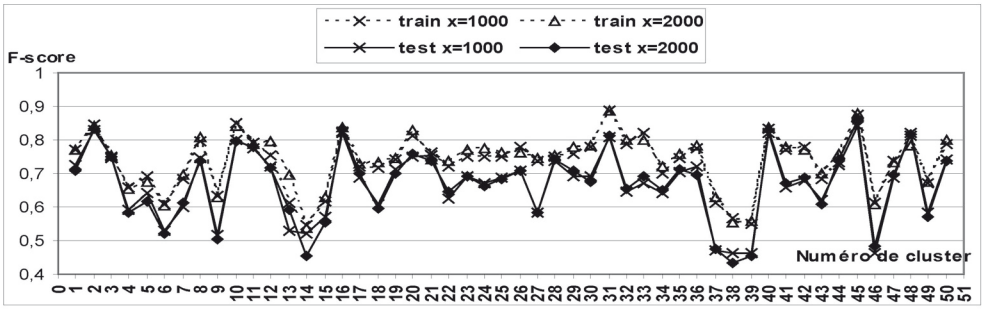

Figure 4 : Evaluation de l'apprentissage des clusters par la liste de k-motifs significatifs

5.4. Comparaison des résultats des deux méthodes

On compare ici avec une autre méthode d'apprentissage des clusters, à base de mots et non de k-motifs, où la phase d'apprentissage (non-supervisé) sur l'ensemble d'entraînement est endogène (méthode des k-means axiales décrite au paragraphe 3.3). Pour l'attribution des classes aux documents dans la phase de validation, l'indice d'Ochiaï est remplacé par la projection des vecteurs-mots sur leur axe de cluster (cf. section 3). La différence entre les deux résultats porte sur la relation entre la performance atteinte (F-score moyen) et le nombre X de motifs ou de termes sélectionnés dans l'ordre des valeurs décroissantes des indicateurs.

Avec la méthode des k-motifs, la performance sur l'ensemble test (dernier dixième du corpus) augmente avec X, se stabilise autour de 2.000 motifs, puis diminue, alors qu'avec celle des « 1-motifs », la performance, plus basse, continue sa croissance quand on augmente le nombre de mots (elle atteindrait 100% avec l'ensemble des mots et des documents). Deux mille k-motifs décrivent donc mieux les clusters, les généralisent mieux, que 5.000 mots : c'est dans la ligne d'un principe général de parcimonie de description. A noter que certains clusters se généralisent mieux que d'autres, comme vu plus haut, et que la description par k-motifs est susceptible de changer leurs contours, dans une mesure que nous n'avons pas établie dans la présente étude.

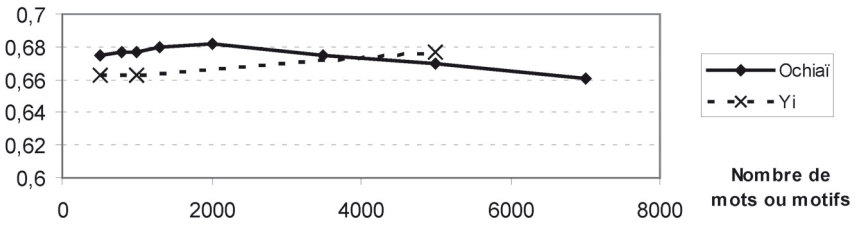

Figure 5 : Taux de reconnaissance sur les 50 clusters, obtenus avec les deux méthodes d'apprentissage

6. Conclusions et perspectives

Nous sommes partis d'un objectif pratique et utilitaire : généraliser une partition, ici obtenue au moyen d'une classification non-supervisée, au moyen d'un ensemble de requêtes le plus réduit possible. Pour ce faire notre exploration s'est appuyée sur des données réelles, de taille suffisante pour utiliser à bon escient la méthodologie de l'apprentissage automatique. Des résultats parfois inattendus ont été établis en chemin :

- Il est techniquement possible d'obtenir par expansion MIDOVA les combinaisons Booléennes de variables binaires caractéristiques d'un corpus d'environ 10.000 documents décrits par 200.000 termes (simples et composés). Ce qui montre qu'en pratique l'explosion combinatoire qu'on pouvait craindre n'a pas lieu, et que les 3-motifs sont les plus nombreux, l'extinction se produisant pour les 6- et 7-motifs.
- Cette expérience a montré que pré-extraire ces termes composés était inutile, et que ceux-ci pouvaient être des *produits* de l'expansion MIDOVA, allégeant d'autant la charge informatique pesant sur l'expansion MIDOVA puisqu'on passerait de 200.000 descripteurs à quelques dizaines de milliers – résultat tout à fait inattendu au départ, mais qui reste à préciser et à approfondir par des études spécifiques.
- Nous avons ensuite réalisé une procédure pour obtenir, au moins sur de très grands ensembles de données binaires, de l'ordre de cent mille individus et plus, 1) des classes non-supervisées stables, 2) la façon de les caractériser de façon compacte en termes de combinaisons Booléennes de variables, interprétables pour une partie d'entre elles comme des termes composés, pour une autre partie comme des associations thématiques à plus longue portée dans les textes. Ce qui confirme, ici dans un cadre non-supervisé, que le pouvoir de généralisation et la compacité de représentation sont indissolublement liés – résultat dont nous avions l'intuition, mais qui n'était nullement acquis à l'avance. On rejoint ainsi le constat paradoxal déjà fait en apprentissage artificiel : on améliore les performances en *augmentant* dans une première phase la taille de l'espace de description des individus, par exemple en utilisant des noyaux polynomiaux, équivalents à des combinaisons Booléennes dans le cas de données binaires, puis en réalisant une séparation linéaire simple dans cet espace augmenté (Guermeur and Paugam-Moisy, 1999). Confirmation aussi du constat de bon sens que les termes composés et expressions figées sont sémantiquement beaucoup plus précis et informatifs que les termes simples.

De nombreuses directions de recherche s'offrent à partir de là, tant dans le domaine pratique – extracteurs de termes, de pseudo-termes, générateurs de requêtes, rétro-requêtage d'un corpus, expansion de requêtes, parallélisation de l'algorithme… – que plus fondamental : compression d'information, apprentissage artificiel, complétion de données manquantes ou de séquences…

Références

Agrawal R. and Srikant H. (1994) Fast algorithms for mining association rules in large databases. In *Research Report RJ 9839*, IBM Almaden Research Center, San Jose, California.

Banerjee A., Dhillon. I., Ghosh J. and Sra S. (2005). Clustering on the Unit Hypersphere using Von Mises-Fisher Distributions. *Journal of Machine Learning Research (JMLR)*, 6: 1345-1382.

Bassecoulard E., Lelu A. and Zitt M. (2007). Mapping nanosciences by citation flows: a preliminary analysis. *Scientometrics*, vol. 70, 3 : 859-880.

Brandes U. and Cornelsen S. (2003). Visual Ranking of Link Structures. *Journal of Graph Algorithms and Applications*, 7 (2) : 181-201.

Cadot M. (2006). *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association*. Doctoral dissertation, University of Besançon, France. Available online at http://www.loria.fr/~cadot/cadot_these_2006.pdf.

Domengès D. and Volle M. (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, 35 : 3-83.

Escofier B. (1978). Analyses factorielles et distances répondant au principe d'équivalence distributionelle. *Revue de Statistique Appliquée*, 26, 4 : 29-37.

Guermeur Y. and Paugam-Moisy H. (1999). Théorie de l'apprentissage de Vapnik et SVM, Support Vector Machines. *Revue Electronique sur l'Apprentissage par les Données (READ)*, vol. 3, 1 : 17-38.

Laurens P., Zitt M. and Bassecoulard E. (2008). Delineation of the field of genomics by hybrid bibliometric method: interaction with experts and validation process. In *10th Int. Conf. on Science & Technology Indicators*, 17-20 September, Wien : Austrian Research Centers GmbH, pp. 323-325. Version étendue à paraître dans *Scientometrics* (2010).

Lebart L. (2007). Which bootstrap for principal axes methods ? In Brito, P., De Carvalho, F., Cucumel, G. and Bertrand, P., editors, *Selected Contributions in Data Analysis and Classification*, Berlin-Heidelberg : Springer, pp. 581-588.

Lelu A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In Diday, E. and Lechevallier, Y., editors, *New Approaches in Classification and Data Analysis*, Berlin : Springer, pp.241-248.

Manly B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo methods in Biology.* Boca Raton, Florida : Chapman & Hall/CRC.

Matusita K. (1955). Decision rules based on distance for problems of fit, two samples and estimation. *Annals of Mathematical Statistics*, vol. 26, 4 : 631-640.

Morineau A., Nakache J.-P. and Krzyzanowski C. (1996). *Le modèle log-linéaire et ses applications. Paris* : Cisia-Ceresta.

Rao C. (1995). A Review of Canonical Coordinates and an Alternative to Correspondence Analysis using Hellinger Distance. *Questiio (Quaderns d'Estadistica i Investigacio Operativa)*, 19 : 23-63.

Renyi A. (1966). *Calcul des probabilités*. Paris : Dunod.

Zitt M., Ramanana-Rahary S. and Bassecoulard E. (2006). Delineating complex scientific fields by a hybrid lexical-citation method: an application to nanosciences. *Information Processing and Management*, vol. 42, 6 : 1513-1531 (erratum Inf. Proc. and Man. 43 (2007) 834).