CASOM : Carte auto-organisée pour l'analyse exploratoire des tableaux de contingence

Rodolphe Priam
IRISA - Projet TexMex
263 av Gén Leclerc F-35000 Rennes
rpriam@gmail.com

Résumé. La visualisation des connaissances par des méthodes d'extraction de l'information pour un corpus de données multimédia est une question pertinente aussi bien en recherche d'information où l'on cherche les meilleurs documents répondant à une requête, qu'en analyse des données où l'on cherche à comprendre quantitativement le contenu du corpus. En effet, en recherche d'information, les corrélations inter-variables permettent d'enrichir la requête de la même façon qu'elles renseignent sur les liaisons interprétables en analyse des données. Une manière générale de répondre à l'objectif posé est l'emploi de méthodes de réduction efficaces qui permettent de mettre en évidence les différentes caractéristiques principales et locales du corpus. Les méthodes de carte auto-organisatrice entrent dans cette optique tout en apportant la dimension supplémentaire d'une carte projective de la distribution et partitionnant le plan en diverses thématiques adjacentes. Ces méthodes rendent appréhendable par l'humain un nuage de points plongé dans un espace de grande dimension par la construction d'une surface discrète qui épouse la forme de sa distribution. Elles offrent ainsi une propriété de cartographie apte à montrer une structure sous-jacente. Dans ce cadre, nous développons une représentation originale des cartes de Kohonen pour des vecteurs textuels bruts. Nous fournissons des indicateurs numériques interprétables et aboutissons à la définition d'une méthode de visualisation synthétique et globale d'un corpus : un *biplot* sous la forme d'un graphe de mots révélant leurs liaisons statistiques, superposable à la projection des documents. La méthode est illustrée par le graphe du vocabulaire extrait d'un corpus de données réelles.

1 Introduction

Dans ce papier, nous introduisons les algorithmes de carte auto-organisatrice et décrivons les principales représentations de cartes auto-organisatrices présentes dans la littérature. Ce cadre posé, nous développons notre méthode CASOM adaptée aux matrices de comptage et mettons en évidence ses diverses propriétés particulières en terme de critère optimisé et métrique. La méthode est illustrée sur un corpus de résumés textuels en construisant des graphes qui montrent les liaisons statistiques entre les termes ou mots du vocabulaire sélectionné dans le corpus ; le graphe de mots a la propriété de se superposer à celui des documents. Cette représentation est également l'occasion d'une réflexion sur la qualité des graphiques résultants. La conclusion dresse le

bilan et propose des perspectives. A la suite de cette introduction, nous présentons notre méthode originale pour une double représentation des lignes et colonnes d'un tableau de contingence, objective, directement interprétable, et à rapprocher de celle de l'Analyse Factorielle des Correspondances. Le principal intérêt d'une telle approche est la visualisation des liaisons entre termes du vocabulaire d'un corpus textuel, simultanément aux thématiques de ce même corpus, contrairement aux méthodes non linéaires classiques qui se contentent de représenter soit l'espace des textes, soit l'espace des termes. En outre, nous sommes en mesure de visualiser l'espace des colonnes d'un tableau de contingence à partir d'une projection des lignes, résultat pertinent pour des matrices fortement asymétriques.

1.1 Introduction aux algorithmes des cartes de Kohonen

L'algorithme original des cartes auto-organisatrices a été introduit par son auteur Kohonen dans les années 1980, d'où le nom de la méthode des cartes de Kohonen [Kohonen, 1997] ou plus généralement *Self-Organizing Maps* (SOM). Un SOM est une méthode de partitionnement d'un ensemble de données multivariées à valeurs réelles, et généralise également l'Analyse en Composantes Principales [Lebart *et al.*, 1997], par la construction d'une surface -principale- discrète[Hastie et Stuetzle, 1989] constituée des centres de la classification. En effet, un algorithme de carte auto-organisatrice est une procédure de classification comparable aux K-means[MacQueen, 1967], mais dans lesquelles une contrainte de voisinage sur les classes est rajoutée. L'algorithme pose un treillis régulier imaginaire souvent rectangulaire, dont chacun des noeuds est affecté à une des classes de la partition à construire. Il utilise cette grille pour lisser les estimations consécutives des centres de classes évalués itérativement à la manière des K-means mais en tenant compte des individus étiquetés comme appartenant à une des classes dont les centres sont placés dans un voisinage -sur le treillis- du centre à adapter, et pondérés par une grandeur décroissante avec leurs distances relatives évaluées sur ce même treillis. L'algorithme construit en fin d'estimation une carte de projection régulière des centres de classe auxquels sont affectées les données les plus proches. Un certain nombre de versions probabilistes de SOM ont été introduites dans la littérature afin de formaliser plus rigoureusement le voisinage flou intervenant dans la pondération du lissage des cartes de Kohonen originales. Ainsi le *Generative Topographic Mapping* ou GTM [Bishop *et al.*, 1998] est un mélange probabiliste de densités gaussiennes dont les centres sont les projections non linéaires des coordonnées bidimensionnelles des noeuds d'une grille régulière sur le plan. Ces méthodes *génératives* sont relativement lourdes en pratique, et surtout peu informatives en analyse des données textuelles où la difficulté de leur estimation numérique les rend difficilement applicables à grande échelle. Nous avons justement récemment développé une variante du SOM que nous nommons CASOM[Priam, 2003] pour *CA by SOM*, adaptée à des vecteurs de comptage et particulièrement aux distributions empiriques textuelles. Elle se rapproche[Priam et Morin, 2001], comme nous allons le voir, de l'Analyse Factorielle des Correspondances ou AFC[Benzecri, 1980], méthode bien connue en analyse quantitative du discours (cf. actes JADT). Il s'agit du même algorithme que le SOM original, muni d'une métrique pertinente pour les données traitées, et induisant des résultats différenciés, notamment au niveau de la visualisation finale. Une visualisation par SOM

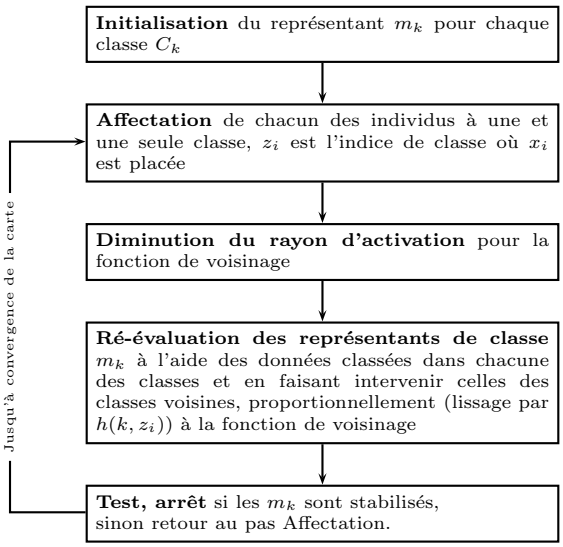

FIG. 1 – Schéma de l'algorithme des Cartes de Kohonen : pour un corpus $\mathcal{D} = \{x_i\}_{i=1}^{i=I}$, un nombre K de classes $\{C_k\}_{k=1}^{k=K}$, et un graphe a priori de voisinage (généralement treillis) sur les classes : (1) chaque classe est représentée par un noeud de coordonnées à deux dimensions ξ_k, (2) une arête est placée entre deux noeuds si les classes correspondantes sont dites voisines, (3) une fonction de voisinage $h(k, \ell)$ décroît avec la distance $||\xi_k - \xi_\ell||^2$ sur le graphe.

s'effectue par post-traitement de la carte de centres obtenus. Les centres construits ne respectent pas forcément les relations de voisinage imaginaire -posé dans l'algorithme- puisque la structure de la distribution multidimensionnelle des données ne se résume pas en général à une *variété* géométrique simple. C'est pourquoi des auteurs proposent dans la littérature diverses représentations, chacune ayant pour objectif de permettre l'interprétation selon un point de vue particulier de la carte auto-organisée.

1.2 Panorama des outils de visualisation par cartes de Kohonen

Etant donné que l'espace de départ -des données vectorielles multidimensionnelles- est souvent de grande dimension, alors que celui d'arrivée -la grille à deux dimensions- est de faible dimension, la projection par SOM induit une certaine déformation apparente. Cette déformation est révélatrice de macro-clusters (ou classes naturelles) séparés par des frontières. Les méthodes de représentation permettent le plus souvent de juger des vraies distances entre les centres comparativement à la grille imaginaire. Selon l'approche préconisée par les auteurs, l'interface résultante revêt des aspects visuels très divers. L'interface *WEBSOM*[Kohonen *et al.*, 2000] montre une véritable image de la distribution projetée du nuage de données par une vue de la matrice U à l'aide d'un

dégradé de couleurs proportionnel aux vraies distances entre classes voisines sur la grille. L'interface *Islands of Musics*[Pampalk *et al.*, 2002] affiche une carte de densité pour des morceaux MP3 de musique, où les îles correspondent aux plus forts taux de remplissage des classes et les étendues d'eau bleutée au vide qui les sépare, L'interface *Hyperbolic SOM*[Ontrup et Ritter, 2001] montre une représentation filaire de la grille pour des résumés de films projetés dans un espace non-euclidien hyperbolique. L'interface *TS-SOM*[Lensu et Koikkalainen, 2002] représente un treillis à hiérarchie pyramidale dans un espace tridimensionnel. L'interface *ET-Map*[Chen *et al.*, 1999] montre la partition induite par une classification hiérarchique [Vesanto et Alhonieni, 2000] des centres de classe d'une carte de Kohonen ; une telle partition permet de clairement révéler les macro-clusters souvent voisins sur la carte, pour ici des pages Internet. Cette représentation est à rapprocher de celle des *treemap*[Bederson *et al.*, 2002]. L'interface *PicSOM*[Laaksonen *et al.*, 1999] montre la sortie brute du résultat auto-organisé de l'interrogation d'une base d'images. De nombreuses autres représentations existent pour notamment projeter les centres de classe et mieux visualiser leurs vraies distances relatives, par divers procédés comme l'analyse en composantes principales, la carte de Sammon ou par interpolation paramétrée. Des auteurs inscrivent parfois également un polyèdre dans une case de la grille où sont placées plus généralement les données étiquetées, polygone dont la taille peut être proportionnelle[Elemento, 1999] au taux de remplissage de la classe. D'autres montrent une représentation graphique des centres vectoriels des classes pour permettre d'apprécier les similitudes entre classes voisines, visualisation à rapprocher de [Inselberg et Dimsdale, 1990], les coordonnées parallèles. Aucune des représentations précédentes ne montre conjointement projection des lignes et colonnes de la matrice des données, exceptée à notre connaissance, une version[Cottrell *et al.*, 1993, Cottrell *et al.*, 2003, Lebart, 2003] du SOM avec métrique du χ^2 qui permet de placer dans les cases de la grille du treillis les modalités des colonnes les plus contributives. La méthode CASOM que nous présentons dans la section suivante permet d'obtenir de véritables *biplots*.

Les cartes auto-organisatrices construisent une partition d'un ensemble de données, donc lorsque ces données sont étiquetées, il devient possible de mesurer un taux d'erreur ou taux de mal classés, après avoir affecté un label de classe pour chacun des noeuds du treillis. On écrit simplement le taux d'erreur empirique de classification comme le rapport du [Nombre de documents bien classés] sur le [Nombre total de documents classés]. Cet indicateur employé sur l'ensemble des données mesure le degré de séparation obtenue par l'algorithme, qui dépend non seulement du pouvoir de séparabilité théorique du modèle qui ne peut être au-dessus du taux de Bayes, mais également du degré de séparation des données considérées. Pour affecter un label à un noeud de la carte, on choisit par exemple l'étiquette représentée majoritairement ou bien celle de la donnée la plus proche du centre. De même si le système est utilisé pour indexer des données, un taux d'erreur similaire[Grossman et Frieder, 1998] est mesurable à partir de la liste des documents retournés en dénombrant le pourcentage empirique parmi les documents retournés qui sont soit pertinents soit hors sujet. D'un point de vue de la visualisation en analyse de données, les étiquettes, lorsqu'elles existent, permettent également de nommer facilement les classes obtenues. Pourtant, en général, les données ne sont pas

étiquetées, et il devient nécessaire de proposer des aides intuitives et quantitatives à l'interprétation. Dans la suite, nous allons définir de tels indicateurs pour la méthode CASOM que nous décrivons dans la section suivante. Une fois la méthode CASOM décrite, nous développerons notre extension à une représentation simultanée des lignes et colonnes d'un tableau de contingence.

2 Méthode CASOM et premières propriétés

Un corpus de I documents est noté $\mathcal{D} = \{x_i\}_{i=1}^{i=I}$ où un document est représenté par un vecteur x_i portant en j-ièmes composantes les fréquences N_{ij} d'occurrences pour les J mots v_j du vocabulaire $\mathcal{V} = \{v_j\}_{j=1}^{j=J}$, vocabulaire restreint[McCallum et Nigam, 1998] défini par sélection de termes ; un document i-ème est une suite de mots "$v_{i1}v_{i2}\cdots v_{i|x_i|}$" à laquelle on associe $x_i = [N_{i1}N_{i2}\cdots N_{iJ}]$, vecteur de comptage. Les K classes C_k de textes ou documents textuels, correspondent aux états k de la v.a. Z qui permet de formaliser la classification par conditionnement probabiliste. Le tableau D des x_i en ligne est appelé généralement *tableau de contingence*, et ici en particulier, *matrice textuelle*. Enfin, $P(\bullet|\theta)$ représente une distribution sur l'espace des mots ou des documents, paramétrée par le vecteur θ. Au maximum de vraisemblance, on note $\theta = \hat{\theta}$, obtenu généralement par algorithme EM[Dempster *et al.*, 1977] (*Expectation-Maximization*) pour les modèles de mélange. On pose $P(x_i|Z = k;\theta)$ la distribution de x_i conditionnellement à l'événement $\{Z = k\}$. Et on écrit $P_{j|k} = P(v_j|Z = k;\theta)$, j-ième composante de $\mathbf{P_k}$, un vecteur J-dimensionnel de probabilités. On note les probabilités empiriques $f_{j|i} = \frac{N_{ij}}{N_{i\bullet}}$, $f_i = \frac{N_{i\bullet}}{N_{\bullet\bullet}}$, $f_j = \frac{N_{\bullet j}}{N_{\bullet\bullet}}$. Enfin, on note $\mathcal{I} = \{1, 2, \cdots, I\}$, $\mathcal{K} = \{1, 2, \cdots, K\}$, $\mathcal{J} = \{1, 2, \cdots, J\}$.

Nous présentons la méthode CASOM pour visualiser des documents et élucidons ses liens avec les méthodes SOM et AFC. On note ξ_k, le vecteur de coordonnées sur la grille à deux dimensions pour la classe de label k. On appelle h, la fonction[1] de voisinage qui prend en compte la topologie de la grille. Par exemple $h(k,\ell) \propto exp(-\frac{1}{2\sigma^2}||\xi_k - \xi_\ell||^2)$ où σ permet de régler la taille du voisinage et donc le degré de lissage. On note la matrice de classification $\mathbf{C}$ de cellules $(c_{ik})_{(i,k)\in\mathcal{I}\times\mathcal{Z}} \in \{0,1\}$, coefficients binaires qui indiquent si le document x_i est dans la k-ième classe ou non. Leur version floue notée $\mu_{ik} \propto h(k, z_i)$, telle que $\mu_{ik} \in [0,1]$ et $\sum_k \mu_{ik} = 1\,\forall i$, tient compte de la fonction de voisinage pour effectuer le lissage spatial sur le treillis. Nous supposons chacun des documents textuels x_i généré selon une loi multinomiale non ordonnée[Minka, 2001] particulière $P(x_i|k,\theta)$ et dont nous estimons les paramètres pour $\sum_j N_{ij}$ supposé fixé, indépendamment pour tout i. On contraint les facteurs multinomiaux à s'"auto-organiser topologiquement" en forçant leur répartition en grille. Ci-après, c_{ik} vaut 1 si $x_i \in C_k$ et 0 sinon ; il s'agit de la logvraisemblance classifiante [Celeux et Govaert, 1992], avec π_k le coefficient mélangeant du k-ième facteur et égal à la proportion d'individus affectés à celui-ci :

$$\mathcal{L}_\mathcal{M}(\theta, \mathbf{C}|\mathcal{D}) = log \prod_{k\in\mathcal{K}} \prod_{i\in\mathcal{I}} \left(\pi_k \prod_{j\in\mathcal{J}} P_{j|k}^{N_{ij}}\right)^{c_{ik}}$$

[1] i.e. $h(k,z) \leq h(\ell,z)$ si $||\xi_k - \xi_z||_{\mathcal{R}^2} \geq ||\xi_\ell - \xi_z||_{\mathcal{R}^2}$ sur la grille.

Initialisation	Initialisation de θ^0, $t = 0$
Pas Affectation	$\forall i\ z_i^t = argmax_{k \in \mathcal{Z}} P(x_i \vert k; \theta^{t-1}) \pi_k^{t-1}$,
Pas Paramètres	$\sigma_t = \eta \sigma_{t-1}$,
	$\forall i, k\ \mu_{ik}^t \propto h(k, z_i^t)$,
Pas Maximisation	$\forall j, k,\ P_{j\vert k}^t = \begin{cases} \frac{\sum_i \mu_{ik}^t N_{ij}+1}{\sum_i \mu_{ik}^t N_{i\bullet}+J} & \text{si } \sum_i \mu_{ik}^t > 0 \\ P_{j\vert k}^{t-1} & \text{sinon.} \end{cases}$,
	$\forall k\ \pi_k^t = \frac{\sum_i \mu_{ik}^t+1}{\sum_i \sum_\ell \mu_{i\ell}^t+K}$,
Test	Si $\vert\vert\theta^t - \theta^{t-1}\vert\vert < \epsilon$ alors $\hat{\theta} = \theta^t$,
	sinon retour au pas Affectation.

FIG. 2 – **Algorithme CASOM**

L'algorithme du TPEM (*Topology Preserving EM*) de [Ambroise et Govaert, 1996] est un algorithme de carte auto-organisatrice qui utilise un mélange de gaussiennes. Les auteurs modifient l'algorithme CEM (*Classification EM*)[Celeux et Govaert, 1992] en rendant flous les c_{ik} binaires en les remplaçant par des coefficients μ_{ik} qui prennent pour valeur une grandeur décroissante avec la distance sur la grille relativement à l'indice fixé z_{ik} qui correspond à l'indice de classe la plus probable pour un document x_i donné. Le TPEM peut également s'interpréter comme une variante de l'algorithme EM, une méthode numérique employée pour la maximisation des vraisemblances à variables latentes telles que posées par les modèles de mélange ; les probabilités a posteriori de classe intervenant dans l'EM y sont remplacées par des valeurs forcées décroissantes autour du noeud le plus probable : le poids μ_{ik} d'un noeud pour un document donné x_i est d'autant plus affaibli qu'il se trouve éloigné (sur la grille) du noeud correspondant à la classe la plus probable pour x_i. Comme le voisinage pris en compte sur la carte dans les interactions entre noeuds est diminué à chaque pas, l'algorithme TPEM se termine[2] en itérations CEM, variante classifiante de l'EM et dont une solution existe, donc la convergence est assurée. Les auteurs justifient la méthode en exhibant une fonction d'énergie minimisée par le SOM (en mode *Batch*) si la taille de l'échantillon est finie. Nous appliquons le même principe pour notre modélisation adaptée au texte en usant d'une loi bien adaptée[McCallum et Nigam, 1998] aux matrices textuelles. Nous utilisons par contre une fonction de voisinage gaussien afin d'effectuer un lissage moins brutal que la fonction en créneau du TPEM. Nous avons également introduit une nouvelle condition pour ne modifier seulement que les centres activés. Nous aboutissons au pas de CASOM qu'il faut répéter[3] jusqu'à la convergence[4]. Le pas de l'algorithme est noté t. On a également noté η un coefficient d'un valeur proche de 1 et permettant de diminuer le rayon de voisinage durant l'apprentissage, par exemple 0,9. Cependant, les justifications de [Ambroise et Govaert, 1996] des liens avec le SOM ne sont plus valables ici puisque nous avons préféré un modèle à variables discrètes. Nous explicitons

[2] On a $h(k, z_i^t) \sim \delta(k, z_i^t)$ pour t grand, et donc $\mu_{ik}^t \sim c_{ik}^t$.

[3] On initialise avec un rayon de voisinage σ_0 dont la taille est celle du coté maximum de la grille. La répartition initiale des documents dans les classes se fait par tirage avec remise.

[4] Voir [Minka, 2001, McCallum et Nigam, 1998] pour le lissage des estimées multinomiales.

le critère minimisé par l'algorithme.

Propriété 2.1
L'algorithme CASOM minimise pour une fonction de voisinage h fixée le critère approché :

$$E_{\mathcal{CASOM}}(\theta, \mathbf{C}|\mathcal{D}) \quad = \quad \sum_{i\in\mathcal{I}} f_i \sum_{k\in\mathcal{K}} c_{ik} \sum_{\ell\in\mathcal{K}} h(k,\ell) \sum_{j\in\mathcal{J}} f_{j|i} \log \frac{f_{j|i}}{P_{j|\ell}}$$

Le critère est pondéré par une probabilité empirique des documents (comme l'AFC) et emploie une distance de Kullback-Liebler entre profils empiriques et centres multinomiaux. En remplaçant par la distance euclidienne classique du SOM, on reconnaît donc le critère minimisé par l'algorithme stochastique du SOM pour une distribution discrète des données. En outre, on peut montrer[Priam, 2003] que la distance est localement une distance du χ^2. On peut montrer également que ce critère est approximativement gaussien sous les hypothèses fortes suivantes : en supposant les données multinomiales et en identifiant les centres estimés avec les vrais centres. Ce résultat demeure peu utile pour les matrices creuses textuelles par rapport à une alternative bayésienne non formulée ici. Par contre, en passant à la limite l'ensemble des tailles de texte, et un chapeau posé sur un paramètre indiquant qu'il s'agit de la solution au maximum de vraisemblance (atteint localement), il vient :

Propriété 2.2
Sous l'hypothèse d'une bonne classification par CASOM, on a pour h fixé et $minN_{i\bullet}$ *grand :*

$$E_{\mathcal{CASOM}}(\hat{\theta}, \hat{\mathbf{C}}|\mathcal{D}) \quad \approx \quad \sum_{k\in\mathcal{K}} (\sum_{i\in\mathcal{I}} \hat{c}_{ik} f_i) \sum_{\ell\in\mathcal{K}} h(k,\ell) \sum_{j\in\mathcal{J}} \hat{P}_{j|k} \log \frac{\hat{P}_{j|k}}{\hat{P}_{j|\ell}}$$

Ce critère permet de retrouver très naturellement un critère connu pour le SOM en modifiant la métrique en une euclidienne. Une pondération supplémentaire apparaît ici naturellement. En outre, des auteurs[de Bodt *et al.*, 2000] ont proposé des tests statistiques par approche *boostrap* pour vérifier la qualité de la répartition des données sur la carte et tester sa pertinence, mais cette méthode exacte reste lourde en pratique à moins de la paralléliser. En alternative, il est possible d'évaluer la qualité de la cartographie en estimant par Monte-Carlo la distribution du critère asymptotique précédent lorsque les centres sont permutés sur la carte.

Nous pouvons finalement énoncer le résultat informel suivant : **CASOM est une généralisation de l'Analyse Factorielle des Correspondances**. En effet, CASOM est un TPEM pour tableau de contingence, or le TPEM s'apparente à une version *Batch* du SOM connu pour généraliser les plans de l'Analyse en Composantes Principales. Nous illustrons l'algorithme sur des résumés[Morin *et al.*, 2000] des publications internes de l'INRIA durant 10 ans et observons la propriété de généralisation en projetant le tableau des $\mathbf{P_k}$, les vecteurs paramètres des lois multinomiales -du mélange contraint de CASOM-, à la manière d'une AFC mais en pondérant plutôt par les coefficients mélangeants $\hat{\pi}_k$. Nous obtenons alors une surface discrète montrant empiriquement (arrêt avant convergence) le lien entre l'AFC et CASOM : une surface

non linéaire résumant l'information contenue dans les plans de l'AFC des données. Nous expliquons dans la section suivante la propriété de représentation particulière de CASOM, très apparentée à celle de la double représentation de l'AFC.

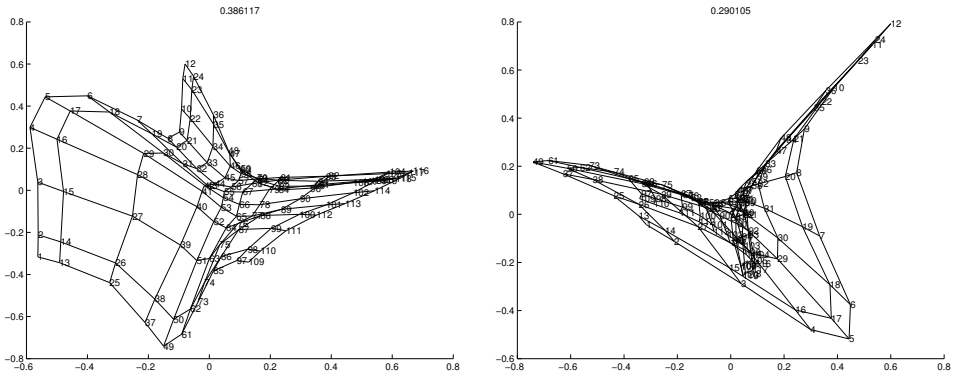

FIG. 3 – Plans factoriels (1,2) (2,3) et inertie projetée pour les multinomiales. On reconnaît le treillis bidimensionnel sur les projections. Les numéros affichés sur les noeuds repèrent les étiquettes de classe k dans CASOM, et les arêtes montrent les relations de voisinage direct entre classes.

3 Double représentation *centro-barycentrique*

Nous étudions dans cette section la projection simultanée des mots du vocabulaire et des textes du corpus. Pour le GTM, méthode paramétrique, $P(x_i|k,\theta)$ est une loi gaussienne multidimensionnelle, de même que pour le SOM à la fin de l'estimation. Dans le cas de CASOM, les moyennes $\mathbf{P_k}$ sont des vecteurs de probabilités obtenus par normalisation en ligne d'un tableau de contingence rassemblant les moyennes pondérées des vecteurs classés et de composantes $P_{j|k}$ qui correspondent aux probabilités des termes v_j dans les classes C_k de documents.

3.1 Construction du *biplot*

Par Bayes, nous obtenons :

$$\hat{P}(k|x_i) = \frac{\hat{\pi_k}\hat{P}(x_i|k)}{\sum_{\ell\in\mathcal{K}}\hat{\pi_\ell}\hat{P}(x_i|\ell)} \text{ avec } \hat{P}(x_i|k) = \prod_j \hat{P}_{j|k}^{N_{ij}}$$

On en déduit la **projection d'un document** sur la carte :

Définition 3.1
La position moyenne du document x_i considérant les probabilités des classes s'écrit :

$$< \xi|x_i >= \sum_{k\in\mathcal{K}} \xi_k \hat{P}(k|x_i)$$

Cette position[5] a le désavantage d'être sensible à la multiplicité de modes de $P(\bullet|x_i;\hat{\theta})$. Par conséquent, un utilisateur doit idéalement vérifier le nombre et la position des modes avant d'interpréter les positions relatives entre deux projetés sur la carte.

Nous proposons ensuite de normaliser le tableau suivant les colonnes. On obtient alors des distributions bivariées discrètes $\hat{P}_{\bullet|j}$ en se plaçant dans un cadre[6] de grille imaginaire. Une façon de procéder est de renormaliser les nombres $\hat{P}_{j|k}$, ou j-ième composantes des centres de classe en calculant $\tilde{P}_{k|j} = \hat{\pi}_k \hat{P}^{\gamma}_{j|k} / \sum_\ell \hat{\pi}_\ell \hat{P}^{\gamma}_{j|\ell}$. On prend $\gamma = 1$ pour une normalisation classique, et $\gamma \neq 1$ bien choisi afin d'accentuer les modes -ou zones actives- tout en évitant d'obtenir des solutions moyennes trop proches du centre de la carte de projection. On décrit formellement la méthode par l'usage de la formule de Bayes : partant des probabilités d'événement mot $\hat{P}_{j|k}$ dans les classes, on calcule (en omettant le passage à la puissance γ) la distribution bivariée des $\hat{P}_{k|j} \propto \hat{P}_{j|k}$ en normalisant pour k.

Définition 3.2
La distribution cartographique d'un mot v_j (respectivement d'un document x_i) s'identifie à la distribution[7] *des $\hat{P}_{k|j}$ (respectivement des $\hat{P}(k|x_i)$) où sont explicités dans le k-ième label, les numéros de ligne et de colonne sur le treillis.*

Elle offre un moyen aisé d'accéder aux *zones d'influence* du mot v_j sur le treillis, et d'observer visuellement ses modes. La distribution cartographique d'un mot s'identifie à la distribution cartographique d'un document qualifié ici de **supplémentaire** et dont la représentation vectorielle s'écrit $x_{[j]} = (0, 0, \cdots, 0, \gamma, 0, \cdots, 0)$ avec pour seule valeur non nulle la j-ième composante. On en déduit également la **projection d'un mot** sur la carte :

Définition 3.3
La position moyenne du mot v_j considérant les probabilités des classes s'écrit :

$$< \xi|v_j >= \sum_{k \in \mathcal{K}} \xi_k \hat{P}_{k|j}$$

Les deux projections précédentes pour les mots v_j en la position $< \xi|v_j >$ pour tout $j \in \mathcal{J}$ et pour les documents x_i en la position $< \xi|x_i >$ pour tout $i \in \mathcal{I}$ se superposent sur la même carte. Cette double représentation s'interprète également en fonction des proximités avec les modes des distributions cartographiques de mot ou de document, mais également avec les noeuds du treillis régulier. Il devient également possible de comparer qualitativement deux mots en confrontant visuellement leur carte de distribution, ou de même mesurer quantitativement une distance d'écart entre les distributions, qui s'écrit par exemple :

[5] Cette projection s'utilise directement avec le TPEM[Ambroise et Govaert, 1996] d'Ambroise de la même manière que notre version pour tableau de contingence. Il peut être préféré une projection à l'aide des seuls voisins de la classe, ou sur un cercle autour du centre le plus proche, dans la direction de la moyenne par exemple.

[6] Ce vecteur de probabilité discrète de composantes $\hat{P}_{k|j}$, peut être considéré comme définissant une loi jointe discrète bivariée croisant les deux variables des modalités ligne et colonne du treillis.

[7] Cette distribution est bivariée lorsque l'on considère le tableau de classes.

Définition 3.4
Une distance entre deux mots v_{j_1} et v_{j_2} est pour CASOM :

$$D(j_1, j_2) = \sum_{k \in \mathcal{K}} \frac{(\hat{P}_{k|j_1} - \hat{P}_{k|j_2})^2}{\hat{P}_{k|j_1} + \hat{P}_{k|j_2}}$$

Notre distance D prend une valeur d'autant plus faible que les mots v_{j_1} et v_{j_2}, ont des distributions superposables. Nous pouvons également fixer un seuil r_α de décision sur la liaison entre deux mots. Une mesure de distance entre un mot et un document s'écrit $D(j, i)$, de la même manière, en calculant l'écart entre leurs deux distributions cartographiques qu'ils conditionnent et qui ont pour support les états de la grille. De même une distance entre deux documents s'exprime similairement. En alternative, nous pourrions utiliser une distance de χ^2 de distribution qui se normalise par la loi du même nom sous les hypothèses de distribution identique. Cette approche permet d'effectuer un test statistique d'adéquation de loi et d'interpréter r_α comme la borne inférieure de la région critique correspondante et de risque α. Enfin, nous nous intéressons maintenant au cas des distributions peu informatives.

Il est clair que plus le sens d'un mot est *multi-thématique*, plus sa *distribution cartographique* devient équidistribuée ; inversement un mot employé le plus souvent dans un seul thème, et donc une seule zone de la carte, aura un mode unique. Une manière de quantifier ce phénomène est l'entropie de la loi discrète $\hat{P}_{\bullet|j}$ qui est d'autant plus faible que la loi est déterministe, donc portée par un seul état. On définit donc :

Définition 3.5
L'entropie cartographique du mot v_j s'écrit comme l'Entropie de Shannon de $\hat{P}_{\bullet|j}$:

$$EC(j) = -\sum_{k \in \mathcal{K}} \hat{P}_{k|j} log_2 \hat{P}_{k|j}$$

On dispose du moyen de découvrir des zones d'intérêt d'un mot au niveau de la carte, en terme de probabilité, résultat inexistant pour le SOM qui estime des centres dans R^J, et de quantifier le degré d'intérêt d'un mot du point de vue de l'information contenue sur la grille, ainsi que du moyen de mesurer des distances mutuelles reflétant la cartographie obtenue. Finalement, une représentation synthétique de ces divers indicateurs consiste à ne choisir que les mots ou documents dont la projection est interprétable, ce qui est possible en sélectionnant les mieux expliqués par CASOM, soit ceux de basse entropie. D'où :

Définition 3.6
Une représentation par graphe du tableau se construit en reliant les projections des mots v_j ou documents x_i par des arêtes lorsque leurs distributions cartographiques sont suffisamment similaires au sens que $D(\bullet, \bullet) \leq r_\alpha$.

Cette visualisation conduit à un graphe synthétique où lignes et colonnes sont affichées simultanément. Il est possible de flécher le graphe en employant une pseudo-distance asymétrique comme celle de Kullback-Liebler au lieu de la $D(\bullet, \bullet)$. Il est également

possible de mettre en évidence une échelle de liaison en choisissant des épaisseurs de flèche variées et proportionnelles à $D(\bullet,\bullet)$. L'objectif des méthodes de carte auto-organisatrice en extraction des connaissances textuelles est souvent d'obtenir un moyen de navigation aisée dans un corpus. Nous quantifions finalement le degré de *pertinence* des distributions cartographiques conservées pour le graphe, et ainsi que celui de leur *interprétativité*.

3.2 Indicateurs d'aide à l'interprétation des projections

En recherche d'information, les corrélations entre termes du vocabulaire d'un corpus permet d'affiner son exploration par requête mot en s'affranchissant d'un éventuel étiquetage en classes. La double représentation illustre ce principe mais nécessite un complément en terme d'indicateurs.

Définition 3.7
Des indicateurs pour l'aide à l'interprétation de la représentation d'un mot v_j par une carte CASOM s'écrivent :

Variance de position	$< \xi^2 \mid v_j >$	$=$	$\sum_k \|\|\xi_k - < \xi \mid v_j > \|\|^2 \hat{P}_{k\mid j}$
Entropie cartographique	$EC(j)$	$=$	$-\sum_k \hat{P}_{k\mid j} log_2 \hat{P}_{k\mid j}$
χ^2 *d'indépendance*	$\chi^2(j)$	$=$	$\sum_v \sum_w \frac{\left(\hat{N}_{(k_{ab}j)} - \frac{\hat{N}_{(k_{a\bullet}j)} \times \hat{N}_{k_{\bullet b}j}}{\hat{N}_{(k_{\bullet\bullet}j)}}\right)^2}{\frac{\hat{N}_{(k_{a\bullet}j)} \times \hat{N}_{(k_{\bullet b}j)}}{\hat{N}_{(k_{\bullet\bullet}j)}}}$
Coefficient de Moran	$I(j)$	$=$	$\frac{K}{2a} \frac{\sum_k \sum_\ell h_{k\ell} (\hat{P}_{jk} - \overline{\hat{P}_{\bullet\mid j}})(\hat{P}_{j\ell} - \overline{\hat{P}_{\bullet\mid j}})}{\sum_k (\hat{P}_{jk} - \overline{\hat{P}_{\bullet\mid j}})^2}$

Les indicateurs pour les documents x_i se définissent similairement. L'entropie cartographique a l'intérêt d'évaluer le désordre statistique d'une carte, mais elle ne prend pas en compte la dimension spatiale qui nous intéresse. C'est pourquoi, nous proposons trois autres indicateurs en complément. Le premier, la *variance de la position* du mot considéré, permet encore de juger de l'importance de l'erreur d'interprétation de la valeur moyenne. La variance ici calculée en module afin de faciliter sa lecture, peut se centrer et réduire au moins par simulation. Le second, le χ^2 *d'indépendance* de la distribution cartographique du mot évalue une mesure de liaison des v.a. ligne-colonne que l'on code par la notation k_{ab} où $\xi_k = (a, b)$ est le vecteur des coordonnées cartésiennes de C_k sur le treillis, ici assimilées aux modalités d'une variable aléatoire bivariée ; les comptages $\hat{N}_{k_{ab}j}$ pour un mot v_j correspondent au tableau des classes non normalisé, et dont les marginalisations à j fixé, suivant une structure de grille des colonnes (classes) sont directement portées sur k pour alléger l'écriture. La loi de $\chi^2(j)$ est connue sous hypothèse d'indépendance[Agresti, 1990] et permet d'exprimer un indicateur normalisé. Le troisième, une *mesure d'auto-corrélation spatiale*, où $\overline{\hat{P}_{\bullet\mid j}}$ est une moyenne empirique des $\hat{P}_{k\mid j}$ supposés continus, permet de quantifier le degré de cohérence de la structure de la carte : le coefficient de Moran[Cliff et Ord, 1981] est un indicateur de l'auto-corrélation spatiale ; on a noté a, le nombre d'arêtes du graphe formé par la grille imaginaire de la carte auto-organisatrice. Il est possible de ramener le coefficient de Moran dans $[0, 1]$ en calculant une "*P-value*" par simulation

par exemple puisque ici les résultats de loi connus dans le cas gaussien ne sont pas applicables. De tels indicateurs nous renseignent au sujet des classes construites par les différents points de vue de la *variance expliquée*, de la *quantité d'information*, de la *corrélation spatiale*, et de la *liaison statistique*. Ces indicateurs entrent naturellement dans le cadre de la méthode CASOM et nous offrent une aide à l'interprétation de la position des mots projetés. La section suivante conclut l'article en mettant en oeuvre la méthodologie précédente et en présentant des développements ultérieurs.

4 Illustration de la méthode et perspectives futures

Notre méthode de visualisation d'un tableau de contingence offre une forme de *feedback* visuel pour enrichir une requête mot afin de découvrir les zones de la carte couvrant les différentes thèmatiques ayant rapport à l'interrogation. La notion de projection globale sur le plan s'apparente aux cartes de Sammon[Sammon, 1969] et celle de *biplot* à la la double représentation barycentrique bien connue en Analyse Factorielle des Correspondances. Un utilisateur navigue dans le graphe en parcourant sur la carte les zones pertinentes pour son interrogation, et en y trouvant le vocabulaire contextuel qui les caractérisent.

4.1 Exemples de graphe pour un corpus textuel

Nous étudions en illustration un corpus de résumés des publications internes de l'IRISA qui ont déjà été interprétées[Morin *et al.*, 2000] à l'aide de l'AFC. Nous formons une carte de taille 12×10, pour ces 1960 textes et 470 mots conservés, puis construisons les sous-graphes des termes *GRAPH*, *GRAPHS*, *INTERFACE* et *KNOWLEDGE*. Seuls les sous-graphes et les distributions cartographiques -tracés sous matlab- sont montrés ici. Les graphiques obtenus par CASOM s'interprétent de la manière suivante. Aux coordonnées d'abcisse et d'ordonnée entières correspond un noeud du treillis, ou classe de textes. Le graphe de mots qui est dessiné sur cette grille, permet donc de caractériser les classes voisines au sens des projections moyennes définies précédemment. Les termes sont affichés suivis de leur fréquence totale. Le mot choisi pour l'étude ou l'interrogation apparaît au centre d'un graphe en étoile, le sous graphe autour du mot requête qui est rappelé en haut de chaque graphique. Les noeuds aux extrémités des branches de l'étoile correspondent aux mots les plus proches du mot requête sélectionné, en terme de distribution cartographique. Les arêtes renvoient à la liaison entre termes du vocabulaire du point de vue de la distance $D(\bullet, \bullet)$ pour le graphe complet. Ici ces arêtes sont rendues inutiles en raison de la visualisation locale proposée, bien qu'elles montrent les distances sur le plan. Ainsi, le terme *KNOWLEDGE* apparaît principalement dans la partie supérieure de la carte construite par CASOM avec deux modes principaux voisins sur la grille. En effet, nous affichons la distribution cartographique du mot *KNOWLEDGE* sous la forme de lignes de niveaux afin d'observer les zones fréquentes du terme sur la carte conjointement au vocabulaire employé. Nous sommes en mesure d'énumérer, en lisant le schéma, les termes du vocabulaire qui se dessinent comme les plus liés à la notion de connaissance, i.e. *DESIGN, INTERFACE, USER, CONCEPTS, REFERENCE, CLASSES, CRITERIA*.

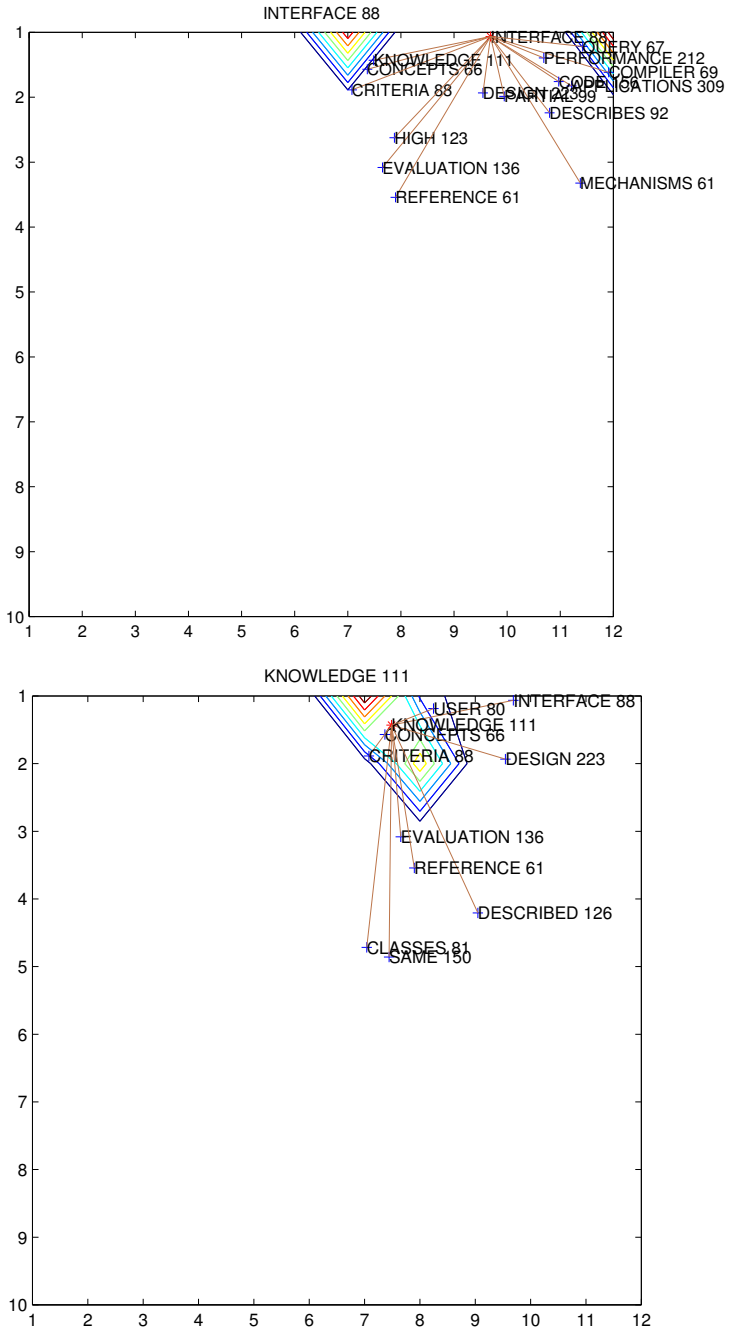

FIG. 4 – Exemples de sous-graphe des mots KNOWLEDGE et INTERFACE

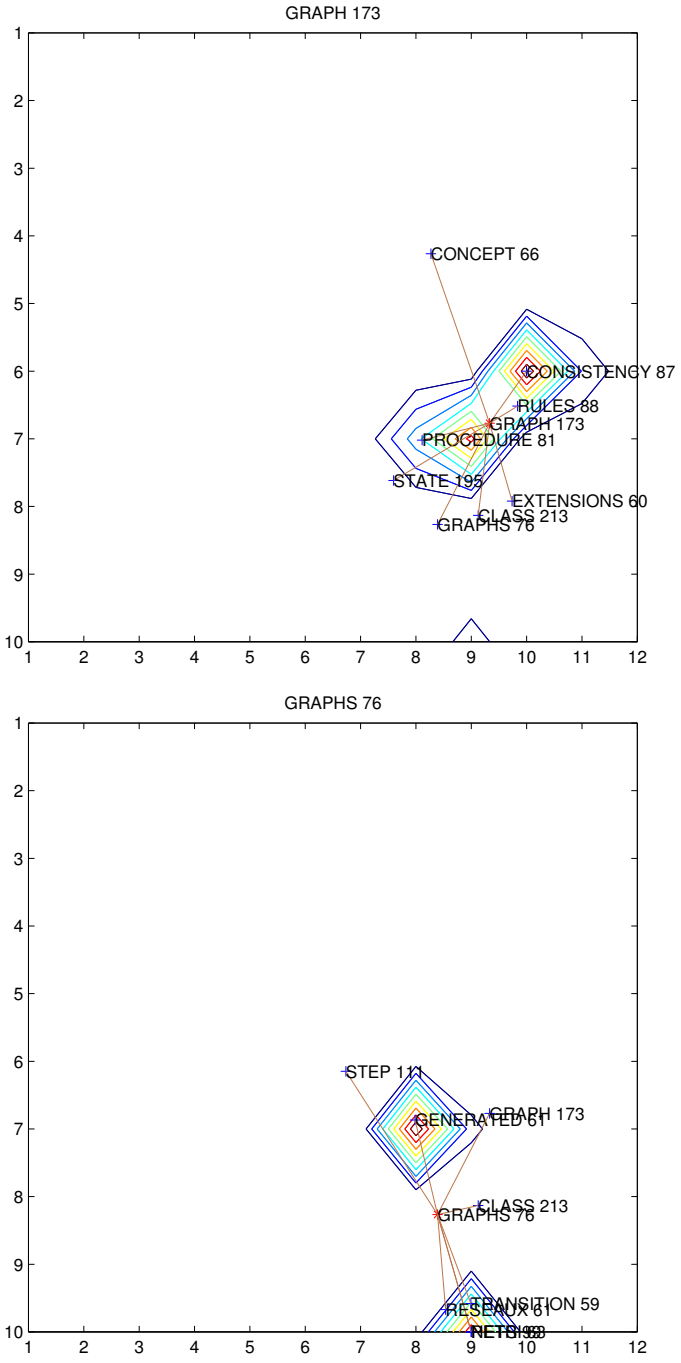

FIG. 5 – Exemples de sous-graphe des mots GRAPH et GRAPHS

Le deux modes directement voisins tendent à montrer l'existence d'un unique théme prépondérant et bien tranché. Le terme *INTERFACE* nous apparaît par contre employé dans deux thèmes bien distincts comme le montre la présence des deux modes éloignés. De même les termes *GRAPH* et *GRAPHS*, l'un au singulier et l'autre au pluriel apparaissent employés dans des documents très différents puisque leur distribution cartographique se superposent mal. Pour valider toutes ces hypothèses, une classification hiérarchique des noeuds de la grille permet de vérifier si les classes correspondantes sont bien éloignées dans l'espace des données. D'autres méthodes de post-traitement telles que le calcul des distances entre centres ou bien leur analyse factorielle sont également envisageables afin d'éviter un retour systématique aux textes originaux.

4.2 Conclusion

Nous avons rappelé comment représenter une carte auto-organisatrice, puis présenté une façon originale de construire un graphe synthétique superposable à la cartographie d'un corpus de textes par une analogie informelle avec la propriété de double représentation barycentrique de l'AFC. Notre méthode CASOM se présente comme une forme synthétique de l'AFC dont elle partage certaines propriétés intéressantes, témoignant de la cohérence de notre approche. Les critères numériques proposés permettent de juger par différents points de vue le contenu d'une carte de distribution bivariée du vocabulaire. L'introduction des statistiques spatiales en qualité de la représentation est nouvelle à notre connaissance. Finalement, nous qualifions de transversale l'approche originale de l'étude des distributions de mots dans CASOM, et que nous mettons en opposition aux approches classiques que nous nommons longitudinales et qui travaillent sur les vecteurs de classes de façon plus globale. De futurs travaux devraient porter sur l'exploitation de la dissociation des effets variables des composantes vectorielles sur la projection. En autres perspectives, une bonne initialisation de l'algorithme CASOM, point crucial, est envisageable par diverses manières comme l'emploi d'un algorithme de SOM robuste sur le tableau réduit afin de réaliser à la suite le calcul des centres dans l'espace du tableau de contingence grâce aux probabilités a posteriori résultantes et éventuellement réduites aux coefficients binaires de classe. Des travaux en cours portent notamment sur l'élaboration d'une double représentation pour le SOM original ainsi que sur la simulation des distributions de nos indicateurs.

Références

[Agresti, 1990] Alan Agresti. *Categorial Data Analysis*. Wiley Series in probability and mathematical statitics, 1990.

[Ambroise et Govaert, 1996] C. Ambroise et G. Govaert. Constrained clustering and kohonen self-organizing maps. *Journal of Classification*, 13(2) :299–313, 1996.

[Bederson *et al.*, 2002] B.B. Bederson, B. Shneiderman, et M. Wattenberg. Ordered and quantum treemaps : Making effective use of 2d space to display hierarchies. *ACM Transactions on Graphics (TOG)*, 4(21) :833–854, 2002.

[Benzecri, 1980] J. P. Benzecri. *L'analyse des données tome 1 et 2 : l'analyse des correspondances*. Paris :Dunod, 1980.

[Bishop *et al.*, 1998] Christopher M. Bishop, Markus Svensén, et Christopher K. I. Williams. GTM : The generative topographic mapping. *Neural Computation*, 10 :215–234, 1998.

[Celeux et Govaert, 1992] G. Celeux et G. Govaert. A classification EM algorithm for clustering and two stochastics versions. *Computational Statistics and Data Analysis*, 14 :315–332, 1992.

[Chen *et al.*, 1999] H. Chen, C. Schuffels, et R. Orwig. Internet categorization and search : A self-organizing approach. *Journal of Visual Communication and Image Representation*, 7 :88–102, 1999.

[Cliff et Ord, 1981] A.D. Cliff et J.K. Ord. *Spatial processes : Models and Application*. Pion Press, 1981.

[Cottrell *et al.*, 1993] M. Cottrell, P. Letremy, et E. Roy. Analyzing a contingency table with kohonen maps : a factorial correspondence analysis. In A.Prieto J.Cabestany, J.Mary, editor, *Proc. IWANN'93*, Lecture Notes in Computer Science, pages 305–311. Springer-Verlag, 1993.

[Cottrell *et al.*, 2003] M. Cottrell, S. Ibbou S., P. Letrémy P., et Rousset P. Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation. *Journal de la Société Française de Statistique*, 144(4), 2003.

[de Bodt *et al.*, 2000] E. de Bodt, M. Cottrell, et M. Verleysen. Are they really neighbor ? a statistical analysis of the SOM algorithm output. *SAMOS preprint*, 2000.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, et D.B. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Ser. B., 39*, 1977.

[Elemento, 1999] Olivier Elemento. Initialisation, convergence, et validation de cartes topologiques de kohonen. Master's thesis, stage INRIA (Yves Lechevallier), 1999.

[Grossman et Frieder, 1998] David A. Grossman et Ophir Frieder. Information retrieval - algorithm and heuristics. *Kluwer Academic Publishers*, 1998.

[Hastie et Stuetzle, 1989] T. Hastie et W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 :502–516, 1989.

[Inselberg et Dimsdale, 1990] A. Inselberg et B. Dimsdale. Parallel coordinates : a tool for visualizing multi-dimensional geometry. In *VIS 90 : Proceedings of the 1st conference on Visualization '90*. IEEE Computer Society Press, 1990.

[Kohonen *et al.*, 2000] T. Kohonen, S. Kaski, K. Lagus, J. Salojrvi, J. Honkela, et V. Paatero et A. Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11 :574–585, 2000.

[Kohonen, 1997] Teuvo Kohonen. *Self-organizing maps*. Springer, 1997.

[Laaksonen *et al.*, 1999] J. Laaksonen, M. Koskela, et E. Oja. PicSOM—A framework for content-based image database retrieval using self-organizing maps. In *Proc. of 11th Scandinavian Conference on Image Analysis (SCIA'99*, pages 151–156, 1999.

[Lebart *et al.*, 1997] Ludovic Lebart, Alain Morineau, et Marie Piron. *Statistique exploratoire multidimensionnelle*. Dunod, 1997.

[Lebart, 2003] L. Lebart. *Analyse des données*, chapter Analyse des données textuelles. Gérard Govaert, Hermés, 2003.

[Lensu et Koikkalainen, 2002] Anssi Lensu et Pasi Koikkalainen. A parallel implementation of the tree-structured self-organizing map. In *6th International Conference on Applied Parallel Computing*, pages 370–379, 2002.

[MacQueen, 1967] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Stat. and Proba.*, volume 1, pages 281–296, 1967.

[McCallum et Nigam, 1998] Andrew McCallum et Kamal Nigam. A comparaison of event models for naives bayes text classification. In *AAAI-98 Workshop on Learning Text Categorization*, 1998.

[Minka, 2001] Thomas P. Minka. Bayesian inference of a multinomial distribution. tutorial, 2001.

[Morin *et al.*, 2000] A. Morin, M. Kerbaol, et J.Y. Bansard. Etude des résumés en français des rapports de recherche d'un institut d'informatique publiés de 1989 à 1998. In *JADT'2000*, 2000.

[Ontrup et Ritter, 2001] J. Ontrup et H. Ritter. Hyperbolic self-organizing maps for semantic navigation. *Advances in Neural Information Processing Systems 14*, 2001.

[Pampalk *et al.*, 2002] E. Pampalk, A. Rauber, et D. Merkl. Using smoothed data histograms for cluster visualization in self-organizing maps. In *Proceedings of the International Conference on Artifical Neural Networks (ICANN'02)*, Madrid, Spain, August 27-30 2002. Springer.

[Priam et Morin, 2001] Rodolphe Priam et Annie Morin. Visualisation des corpus textuels par Treillis de Multinomiales auto-organisées - Generalisation de l'Analyse Factorielle des Correspondances. *Revue RSTI-ECA (Actes EGC'2002)*, 1(4) :407–412, 2001.

[Priam, 2003] Rodolphe Priam. *Méthodes de carte auto organisatrice par mélange de lois contraintes. Application à l'exploration dans les tableaux de contingence textuels.* PhD thesis, Université de Rennes 1, Octobre 2003.

[Sammon, 1969] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 5(18C) :401–409, may 1969.

[Vesanto et Alhonieni, 2000] Juha Vesanto et Esa Alhonieni. Clustering of the self-organizing map. *IEEE Neural Networks*, 3(11), 2000.

Summary

After an overview of some methods for visualizing a self-organizing map output, we explain how to build a graph of words for a textual corpus using the CASOM method. Our graph has the property to be showable at the same time that the mesh of classes of documents. Quality of the visual datamining method is formally introduced through quantitative indicators.