Un nouvel algorithme incrémental de gaz neuronal croissant basé sur l'étiquetage des clusters par maximisation de vraisemblance : application au clustering des gros corpus de données textuelles hétérogènes

Jean-Charles Lamirel(*), Zied Boulila(*), Maha Ghribi(**), Pascal Cuxac(**), Claire François(**)
jean-charles.lamirel@loria.fr, maha.ghribie@inist.fr, pascal.cuxac@inist.fr, claire.francois@inist.fr

(*)INRIA team TALARIS-LORIA, Vandoeuvre les Nancy, France
(**)INIST-CNRS, Vandoeuvre les Nancy, France

Mots clefs :
Veille scientifique et technologique, clustering, classification incrémentale, gaz neuronal, données textuelles, qualité
Keywords:
Scientific and technical observation, clustering, incremental classification, neural gas, textual data, quality
Palabras clave :
Escudriñar científico y tecnológico, clustering, clasificación incremental, gas neuronal, datos textuales, calidad

Résumé

Dans le cadre de la veille ou de l'analyse prospective, il est très courant d'avoir recours aux méthodes de clustering pour traiter de gros volumes de données textuelles. Les algorithmes de clustering affichent généralement de bonnes performances dans le cas où les corpus à traiter sont de nature homogène. Cela vaut particulièrement pour les algorithmes de clustering neuronaux, et encore plus spécifiquement pour les récentes versions adaptatives de ces algorithmes, comme l'algorithme incrémental de gaz neuronal croissant (IGNG). Cependant, ce papier illustre clairement la chute drastique de performance de la plupart de ces algorithmes dans le cas plus réaliste où les corpus à traiter s'avèrent être de nature hétérogène, ou polythématique. Dans ce contexte, des mesures spécifiques de qualité de clustering et de nouvelles techniques d'étiquetage des clusters qui sont indépendantes de la méthode de clustering utilisée sont exploitées pour l'évaluation des performances des méthodes. Un nouvel algorithme de gaz neuronal croissant exploitant une mesure de similarité basée sur la maximisation de la qualité de l'étiquetage des clusters est ensuite présenté comme une alternative à l'algorithme IGNG original basé sur la distance euclidienne. Nous montrons que cette solution permet d'obtenir un accroissement très significatif de performance pour le clustering des données textuelles polythématiques. Celle-ci fournit également par ailleurs un véritable caractère incrémental à l'algorithme proposé.

1 Introduction

Les méthodes neuronales de clustering partagent le principe de prendre en considération des relations de voisinage entre les clusters, qu'elles soient prédéfinies (topologie fixe), comme dans le cas des « cartes auto-organisatrices » (SOM) [14], ou dynamiques (topologie libre), comme dans celui des « gaz neuronaux » qu'ils soient statiques (NG) [23], ou croissants (GNG) [6]. Comparativement aux méthodes de clustering usuelles, comme

K-MEANS [25], cette stratégie les rend moins sensibles aux conditions initiales, ce qui représente un avantage déterminant dans le cadre de l'analyse des données textuelles, qui sont souvent représentées comme des données éparses associées à des espaces de description fortement multidimensionnels.

Les avantages et les inconvénients théoriques des différents algorithmes de clustering existants sont une chose, mais qu'en est-il dans la réalité? Une comparaison plus claire du comportement de ces algorithmes utilisant une base commune, et plus particulièrement une estimation fiable de la qualité de leurs résultats sur des données complexes comme les données textuelles hétérogènes ou polythématiques s'avérait nécessaire pour apporter plus de lumière sur leur avantage et sur leurs inconvénients effectifs. Elle s'avérait également indispensable pour décider de l'exploitation ultérieure de certains d'entre eux dans le cadre de la classification incrémentale. Nous avons donc mené une expérimentation exhaustive de ces algorithmes sur deux types de données différents, à savoir des données homogènes et des données polythématiques, le contexte de ces dernières étant celui qui se rapproche le plus, dans le cadre statique, des contraintes intrinsèques de la classification incrémentale.

Nous avons pour cela exploité nos propres mesures génériques d'évaluation de la qualité du clustering qui sont indépendantes de la méthode de clustering et que nous présentons par la suite. Nous exposons également ci-après une comparaison du comportement des méthodes de clustering neuronales que nous avons présentées sur des données homogènes et hétérogènes en utilisant une méthode de clustering non neuronale, en l'occurrence la méthode Walktrap, comme méthode de référence. Nous présentons finalement les améliorations que nous avons développées à la fois pour obtenir des résultats satisfaisants sur des données complexes, comme les données polythématiques, et parallèlement, pour obtenir des méthodes offrant un comportement réellement incrémental.

2 Les méthodes de classification dérivées de SOM

L'approche SOM est basée sur l'observation biologique que le cortex possède la capacité particulière de réduire la représentation de grands ensembles de données sous la forme de cartes auto-organisées. Comme le montre Kohonen [14], une cartographie d'un espace de données multidimensionnel sur une grille bidimensionnelle de neurones qui regrouperont après apprentissage les propriétés synthétiques des données considérées peut ainsi être définie. L'algorithme d'apprentissage de SOM est présenté en détails dans [14]. Il se compose de deux procédures de base appliquées à chaque nouvelle donnée d'entrée : (1) choix d'un neurone gagnant pour la donnée sur la grille, et, (2) mise à jour des poids, ou composantes, associés au vecteur de référence du neurone gagnant et de ceux de ses neurones voisins en fonction de cette donnée. Chaque neurone pouvant être assimilé à un cluster, l'algorithme SOM peut donc être considéré comme un algorithme de clustering du type « le gagnant emporte le plus »[1]. Une fois l'apprentissage achevé, les données d'apprentissage peuvent être affectées individuellement aux neurones, ou clusters, de la grille.

Les cartes SOM ont été employées avec succès pour de nombreuses applications du domaine général de l'analyse de données textuelles, comme pour le clustering de comptes-rendus de réunion ou celui de données socio-économiques [29], ou encore pour la cartographie et le feuilletage des fonds documentaires [16]. Kaski et al. ont notamment mis au point une adaptation spécifique de SOM, appelée WEBSOM, pour l'analyse des grands fonds de documents, en exploitant des techniques de projection aléatoires visant à réduire la dimension descriptive des documents [12]. La méthode MultiSOM a introduit de son coté la communication entre plusieurs cartes SOM associées à des points de vue différents définis sur les

[1] Contrairement à K-MEANS, qui peut être considéré comme un algorithme de type « le gagnant remporte le tout », du fait que chaque donnée d'entrée n'influence qu'un seul cluster à la fois au cours de l'apprentissage.

mêmes données [17], ce qui revient à permettre de croiser de manière non supervisée des analyses basées sur des critères multiples. Cette dernière approche a été appliquée avec succès à l'analyse des notices de brevets [18], ainsi qu'à celle des données Web ([7],[20]). Finalement, des versions incrémentales de SOM intégrant des processus de croissance hiérarchique de la grille permettent de suivre l'évolution des caractéristiques des données [24]. L'avantage général de l'approche SOM est qu'elle combine de manière homogène dans un même modèle un processus de clustering avec un processus complémentaire de projection des résultats. Cependant, le fait que les cartes SOM représentent des structures discontinues rend l'approche spécialement sensible aux données marginales [15]. Le défaut principal de cette méthode reste malgré tout celui de ne pas permettre représenter très fidèlement les distributions de données complexes en raison de la structure topologique fixe de la grille.

Au contraire, l'algorithme « Neural Gas » (NG) [23], utilise un processus dynamique d'estimation du voisinage des neurones vis-à-vis de chaque donnée d'entrée. Les changements de poids ne sont pas déterminés par les distances relatives entre les neurones dans un treillis typologiquement pré-structuré, mais par la distance relative entre ceux-ci dans l'espace d'entrée, d'où la désignation de « gaz neuronal » attribuée à ce type de réseau. De fait, grâce à la relaxation des contraintes topographiques par rapport à SOM, NG tend à mieux représenter la structure des distributions de données complexes, menant théoriquement à de meilleurs résultats de classification. Néanmoins, un des inconvénients principaux de cet algorithme est que le nombre de neurones, matérialisant le nombre final de clusters, représente un paramètre fixe.

L'algorithme « Growing Neural Gas » (GNG) [6] résout le caractère statique de l'algorithme NG en mettant en avant le concept du réseau évolutif. En effet, dans cette approche, le nombre de neurones est adapté pendant la phase d'apprentissage en fonction des caractéristiques de la distribution de données analysée. GNG donne ainsi la possibilité de créer et de supprimer des neurones, ainsi que des connexions structurelles de voisinage entre ces derniers[2]. Le processus de suppression se fonde sur la notion d'âge limite, d'une connexion. Au temps t, si une connexion atteint cet âge limite, sans avoir été renouvelée, ou rafraichie, elle est automatiquement supprimée; les neurones isolés sont ensuite supprimés en conséquence. D'autre part, la création des neurones est opérée seulement périodiquement (à chaque T itérations ou méta-période de temps) entre les deux neurones voisins qui ont cumulés l'erreur la plus importante pour représenter les données. Différents critères d'arrêt peuvent être employés, comme un nombre maximal de neurones ou un changement minimal entre les vecteurs de référence des neurones entre deux périodes de temps. Cependant, ces critères peuvent dans certains cas ne pas être atteints, particulièrement avec des données multidimensionnelles complexes ou éparses. Ceci représente un inconvénient important de la méthode GNG.

L'algorithme « Incremental Growing Neural Gas » (IGNG) [26] représente une adaptation de l'algorithme GNG qui relaxe la contrainte d'évolution périodique du réseau. Par conséquent, dans cet algorithme, un nouveau neurone est créé chaque fois que la distance de la donnée d'entrée courante aux neurones existants est supérieure à un seuil préfixé σ . La valeur est un paramètre global qui correspond à la distance moyenne des données d'entrée vis-à-vis du centre de leur distribution. Par ailleurs, chaque nouveau neurone passe par une phase embryonnaire durant laquelle il ne peut pas être supprimé, même s'il a perdu toutes ses connexions de voisinage. La phase embryonnaire est un paramètre fixe qui correspond à un nombre donné d'étapes de temps après lesquelles un neurone devient mature. Prudent et Ennaji arguent du fait que cette phase permet de réduire l'impact des données bruitées durant l'apprentissage. A l'issue de la phase d'apprentissage, les données sont projetées uniquement sur les neurones matures. Prudent et Ennaji ont montré que la méthode IGNG produisait de meilleurs résultats que les méthodes neuronales concurrentes sur les

[2] Ces connexions, dites « hebbiennes » [10] ou compétitives, sont crées au cours de l'apprentissage entre le neurone gagnant et son concurrent direct. Elles permettent de structurer dynamiquement le réseau créé, et peuvent être exploitées dans tous les algorithmes neuronaux à topologie libre.

distributions-test usuelles. Cependant, le fait que le paramètre σ doive être calculé avant l'apprentissage et dépende globalement de l'ensemble des données d'entrée ne plaide pas en faveur du caractère incrémental de la méthode, même si celui-ci a été défendu par ses auteurs.

L'objectif principal de l'algorithme « Improved Incremental Growing Neural Gas » (I^2GNG) [9] est de résoudre la faiblesse de l'algorithme IGNG en termes de comportement incrémental. Dans ce but, l'algorithme I^2GNG exploite une valeur variable du seuil σ qui est calculée à chaque étape de l'apprentissage et qui dépend de chaque neurone (c.-à-d. de chaque cluster). Les données sont dynamiquement associées aux neurones durant l'apprentissage. Par conséquent, au temps t, le seuil σ de chaque neurone correspond à la distance moyenne de son vecteur de référence à ses données actuellement associées. L'algorithme I^2GNG a été testé avec succès pour la classification non supervisée de documents de facturation. Cependant, l'inconvénient principal de celui-ci semble être lié à son initialisation. En effet, une valeur de seuil par défaut doit être spécifiée à la création de chaque nouveau neurone, et aucune solution précise n'a été proposée jusqu'ici par les auteurs pour traiter correctement ce problème.

3. Evaluation des résultats de clustering basée sur les notions de précision et de rappel

En classification non supervisée (clustering), le fait de ne pas avoir de classification de référence sur laquelle s'appuyer représente un lourd handicap pour évaluer la performance des algorithmes. Il existe certes des indices de qualité basés sur des calculs de distance dont les plus connus sont les inerties intra-classe et inter-classes [22] :

- L'inertie intra-classe permet de mesurer le degré d'homogénéité entre les données associées à une classe. Elle calcule leurs distances par rapport au point représentant le profil de la classe.

- L'inertie inter-classes mesure le degré d'hétérogénéité entre les classes. Elle calcule les distances entre les points représentant les profils des différentes classes de la partition.

Parmi les autres indices de qualité utilisant la distance entre individus, ou données, qui ont été développés, l'on citera l'indice de Dunn [5], l'indice de validation de Davies-Bouldin [2] et la Silhouette [28]. Tous ces indices, ne permettent cependant pas d'estimer la qualité du clustering dans bon nombre de cas, comme dans celui des données textuelles [8], [13]. Dès 2004, nous avons développé une approche alternative basée sur des mesures de Rappel, Précision et F-mesure non supervisée exploitant les descripteurs des données associées aux classes [19], que nous avons affinée depuis.

Dans le domaine de la recherche d'information, le Rappel *R* représente le rapport entre le nombre de documents pertinents restitués pour une requête donnée et le nombre total de documents pertinents qui auraient été trouvés dans la base de données documentaire. La Précision *P* représente le rapport entre le nombre de documents pertinents qui ont été restitués pour une requête donnée et le nombre total de documents retournés pour ladite requête. Le Rappel et la Précision se comportent en général de manière antagoniste : quand le Rappel augmente, la Précision baisse, et inversement. La fonction *F* a ainsi été mise en place pour identifier le meilleur compromis entre le Rappel et la Précision. Elle est représentée par la moyenne harmonique :

$$F = \frac{2(R \times P)}{R + P} \quad \text{(Eq. 1)}$$

Le principe des mesures de Rappel et de Précision adapté au cas de la classification non supervisée est illustré à la figure 1.

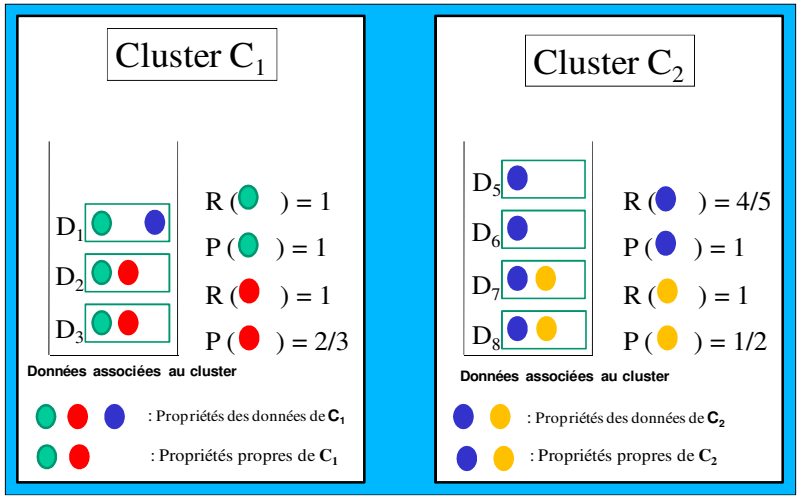

Figure 1. Principe des indices de Rappel(R)-Précision(P) non supervisés.

Le Rappel permet de mesurer l'exhaustivité du contenu des clusters, lié à la présence de propriétés qui leur sont spécifiques (que nous nommons « propriétés propres »). Plus un cluster présente un ensemble de propriétés propres qui lui sont exclusives, plus il se distingue des autres clusters, et donc plus le critère d'hétérogénéité entre clusters est renforcé.

La Précision mesure l'homogénéité des clusters en termes de proportion de données contenant les propriétés propres de ces premiers. Plus les données associées à un cluster présentent des propriétés propres communes, plus elles sont similaires entre elles, et donc plus le critère d'homogénéité à l'intérieur des clusters est renforcé.

Plus précisément, soit une partition $P = (C_1;\ldots;C_k)$ issue d'une classification non supervisée d'un ensemble de documents. Pour un cluster C, nous pouvons définir l'ensemble des propriétés propres S_c :

$$S_C = \left\{ p \in d, d \in C_i \, C \mid \overline{W}_C^p = \max_{C' \in P} (\overline{W}_{c'}^p) \right\}, \text{avec } \overline{W}_C^p = \frac{\sum_{d \in C} W_d^p}{\sum_{C' \in P} \sum_{d \in C'} W_d^p} \quad \text{(Eq. 2)}$$

où W_p^d représente le poids de la propriété *p* pour un document *d.*

et $\overline{W}_C^p$ représente le rapport du poids cumulé de la propriété *p* dans le cluster *C* par rapport à son poids total dans la partition. Les propriétés propres maximisent donc le rappel pondéré ($\overline{W}_C^p$).

Nous pouvons donc définir l'ensemble des clusters propres de la partition *P* comme suit :

$$\overline{P} = \left\{ C \in P \middle| S_C \neq \varnothing \right\} \quad \text{(Eq. 3)}$$

Dans cet ensemble de clusters propres, nous définissons alors les Macro Rappel- Précision comme les valeurs moyennes de Rappel et de Précision pour l'ensemble des clusters. Ils prennent les formes suivantes :

$$R_M = \frac{1}{\left|\overline{P}\right|} \sum_{C \in \overline{P}} \frac{1}{\left|S_C\right|} \sum_{p \in S_C} \frac{\left|c_p\right|}{\left|P_p\right|} \; ; \; P_M = \frac{1}{\left|\overline{P}\right|} \sum_{C \in \overline{P}} \frac{1}{\left|S_C\right|} \sum_{p \in S_C} \frac{\left|c_p\right|}{\left|c\right|} \quad \text{(Eq. 4)}$$

où c_p présente l'ensemble des données du cluster *C* possédant la propriété *p* et P_p représente l'ensemble des données de la partition *P* possédant la propriété *p.*

Le Macro-Rappel et la Macro-Précision ont des comportements inverses en fonction du nombre de clusters. Ainsi, ces indices permettent d'estimer de manière globale un nombre optimal de clusters pour une méthode donnée et pour un ensemble de données fixé. La meilleure partition est dans ce cas celle qui minimise l'écart entre leur valeur.

Nous avons préalablement montré [19] qu'un des avantages déterminants de cette approche, qui s'inspire de l'analyse du comportement des classificateurs symboliques, est d'être indépendante de la méthode de clustering utilisée, contrairement aux approches basées sur la distance. Elle permet donc de comparer différentes méthodes entre elles. Cependant, son défaut principal est que la Macro-Précision, en particulier, est peu sensible à la présence de clusters hétérogènes de fort effectif, surtout dans le cas de l'existence conjointe d'un nombre important de clusters de faible taille [8]. En conclusion, même si les Macro-mesures permettent d'estimer le nombre optimal de clusters pour une méthode donnée, elles ne permettent pas pour autant d'estimer la qualité intrinsèque de la partition produite par ladite méthode.

Pour corriger cela, nous définissons ci-après de nouveaux indices de Micro – Rappel/Précision, calculés en moyennant directement les valeurs de Rappel/Précision sur l'ensemble des propriétés propres, et non plus sur les clusters :

$$R_m = \frac{1}{\left|L\right|} \sum_{c \in \overline{C}, p \in S_c} \frac{\left|c_p\right|}{\left|P_p\right|} \; ; \; P_m = \frac{1}{\left|L\right|} \sum_{c \in \overline{C}, p \in S_c} \frac{\left|c_p\right|}{\left|c\right|} \quad \text{(Eq. 5)}$$

où |*L*| représente la dimension de l'espace de description des documents.

Les Micro- Rappel/Précision possèdent des caractéristiques générales analogues aux Macro- Rappel/Précision. Cependant, en les comparant avec ces derniers indices, il devient possible d'identifier des résultats de clustering hétérogènes. En effet, dans ce dernier cas, les Précisions des clusters de petite taille ne compenseront plus celles des clusters de grande taille et les propriétés imprécises présentes dans ces derniers, s'ils s'avèrent

hétérogènes, auront un effet considérable sur la Micro-Précision. Par conséquent, même si la Macro- et la Micro-Précision mesurent toutes deux le degré d'homogénéité des clusters, l'écart entre ces deux mesures permet de confirmer la présence de clusters hétérogènes de taille importante.

D'une manière complémentaire, nous définissons des étiquettes dont le rôle est de mettre pus spécifiquement en lumière les caractéristiques ou les propriétés spécifiques des clusters inhérents à une partition. Cet étiquetage peut être ainsi employé pour visualiser ou synthétiser des résultats de clustering [21], ou bien encore pour valider ou optimiser l'apprentissage d'une méthode de clustering [1]. Il peut se baser sur les propriétés endogènes des données aussi bien que sur leurs propriétés exogènes. Les propriétés endogènes des données représentent celles qui sont employées durant le processus de clustering. Les propriétés exogènes représentent des propriétés complémentaires, ou des propriétés spécifiques de validation. Des mesures de pertinence d'étiquette peuvent être directement dérivées des indices de qualité que nous avons présentés précédemment.

Aussi, pour une propriété *p*, le *Rappel d'étiquette L-R* dérive directement de l'équation 2. Il est exprimé comme :

$$L\_R(p) = \overline{W}_c^p \quad \text{(Eq. 6)}$$

La *Précision d'étiquette L-P* peut être exprimée comme :

$$L\_P(p) = \frac{\sum_{d \in c} W_d^p}{\sum_{p' \in d, d \in c} W_d^p} \quad \text{(Eq. 7)}$$

En conséquence, pour un cluster *c*, l'ensemble d'étiquettes L_C qui peuvent lui être attribuées est l'ensemble des propriétés des données qui maximisent la *F-mesure d'étiquettage* à l'intérieur de ce cluster, c'est-à-dire celle qui vérifient :

$$L_c = \left\{ p \in d, d \in c \mid L\_F(p) = \underset{c' \in C}{Max}(L\_F(p)) \right\} \quad \text{(Eq. 8)}$$

où la *F-mesure d'étiquettage* $L\_F(p)$ peut être définie comme :

$$L\_F(p) = \frac{2(L\_R(p) \times L\_P(p))}{L\_R(p) + L\_P(p)} \quad \text{(Eq. 9)}$$

Dès lors que le Rappel d'étiquette est équivalent à la probabilité conditionnelle *P(c|p)* et la Précision d'étiquette est équivalente à la probabilité conditionnelle *P(p|c)*, cette stratégie d'étiquetage peut être comprise comme une approche de maximisation de vraisemblance, relativement à la définition originale donnée par Dempster et al. [3].

Dans la section suivante, nous utiliserons les mesures de Macro/Micro Précision, Rappel, F-mesure, ainsi que la F-mesure d'étiquettage pour réaliser les évaluations numériques des résultats de clustering. Nous réaliserons également une analyse qualitative par le biais des étiquettes associées aux clusters.

4. Comparaison des méthodes de clustering

Le but de notre première expérience est de comparer le comportement des méthodes de clustering neuronales décrites en section 2 sur deux types différents de corpus de données textuelles, à savoir un corpus thématiquement homogène et un corpus polythématique, de manière à apporter plus de lumière sur les qualités et les défauts effectifs de ces méthodes dans un contexte suffisamment général.

Notre corpus-test de données thématiquement homogènes est un ensemble de 1000 notices de brevets se rapportant à la technologie des huiles moteurs enregistrées durant l'année 1999. Un index brut est produit à partir du texte associé au sous-champ des notices décrivant le domaine d'utilisation des brevets, en exploitant un analyseur lexicographique permettant d'extraire des termes composés [11]. Cet index brut est ensuite normalisé en fusionnant les termes synonymes (par exemple « ingénierie des huiles » et « fabrication des huiles »). Chaque notice de brevet est ensuite indexée par une sélection de termes issue de l'index normalisé. Nous avons finalement appliqué un seuil de fréquence de 2 sur les termes d'indexation, avec comme conséquence de réduire l'espace de description des notices à un espace de 234 mots-clés. Le corpus résultant peut être considéré comme un corpus homogène puisqu'il couvre le champ élémentaire du domaine d'utilisation des brevets avec un vocabulaire normalisé limité et contextuel.

Notre corpus-test de données hétérogène, ou polythématique, représente de son coté un ensemble de 1341 notices bibliographiques issues de la base de données de PASCAL de l'INIST et couvrant 1 année complète de recherche (i.e. 2005), tous domaines confondus, et impliquant au moins un laboratoire de recherche Lorrain[3]. Comme dans le cas des notices de brevets, la structure d'une notice bibliographique est une structure multi-champs (titres, auteurs, affiliations, résumé, mots-clés, …) qui synthétise l'information contenue dans la publication correspondante. Dans notre expérience, seuls sont pris en compte les termes présents dans le champ des mots-clés. Un seuil de fréquence de 3 est finalement appliqué sur ces termes d'indexation, avec comme conséquence de réduire l'espace de description des notices à un espace de 889 mots-clés. Comme ces mots-clés couvrent malgré tout un grand ensemble de sujets différents (aussi éloignés les uns des autres que la médecine, la physique structurelle ou la sylviculture), le corpus résultant peut cette fois être considéré comme un corpus fortement polythématique.

Dans les deux cas, un prétraitement est appliqué aux index résultants des notices de manière à en obtenir une représentation vectorielle. Les vecteurs obtenus sont ensuite pondérés selon le schéma de pondération IDF [27], afin de diminuer l'effet des termes d'index les plus répandus.

Les méthodes de clustering neuronales que nous avons précédemment décrites ont toutes été considérées dans nos expérimentations. Nous avons également considéré la méthode K-MEANS [25] dans ces expérimentations afin de confronter aux méthodes susdites une méthode de référence de la catégorie des méthodes de clustering non neuronales. Pour chaque méthode neuronale, ainsi que pour la méthode K-MEANS, nous avons opéré plusieurs apprentissages différents, en faisant varier le nombre de clusters dans le cas des méthodes statiques (SOM, NG), et les paramètres de

[3] Ces notices ont été extraites de la base à partir de l'interface STANALYST de l'INIST.

voisinage dans le cas des méthodes dynamiques (GNG, IGNG, I2GNG). Nous avons finalement conservé la meilleure partition produite par chaque méthode en nous basant sur l'examen des valeurs de Macro Rappel-Précision définies à la section 2 (Equation 1-4).

Dans le cas de la méthode IGNG, nous avons optimisé le processus en en faisant varier le paramètre de voisinage σ autour de la valeur optimale théorique σ_T. Nous avons observé que les meilleurs résultats de clustering étaient obtenus par l'emploi de valeurs plus petites que σ_T.
Nous avons donc défini une fonction d'évolution du paramètre σ telle que :

$$\sigma_1 = \sigma_T,$$
$$\sigma_{i+1} = \beta \sigma_i \text{ , avec } \beta < 1.$$

Dans le cas d'IGNG, nous avons optimisé le processus en faisant varier σ selon cette dernière stratégie. Le même processus est également opéré avec la méthode I^2GNG en utilisant cependant une variation du paramètre σ relative à chaque cluster en lieu et place d'une variation absolue.

Les résultats de clustering obtenus sur les données thématiquement homogènes et ceux obtenus sur les données polythématiques sont présentés dans le tableau 1.

	DONNEES THEMATIQUEMENT HOMOGENES			**DONNEES POLYTHEMATIQUES**		
METHODE DE CLUSTERING	NBR OPTIMAL DE CLUSTERS	F-MESURE MACRO	F-MESURE MICRO	NBR OPTIMAL DE CLUSTERS	F-MESURE MACRO	F-MESURE MICRO
K-MEANS	158	0.89	0.62	155	0.88	0.03
SOM	177	0.78	0.66	**289**	**0.47**	**0.40**
NG	160	0.85	0.86	160	0.59	0.33
GNG	170	0.80	0.80	--	--	--
IGNG	**92**	**0.90**	**0.85**	378	0.58	0.21
I^2GNG	32	0.58	0.38	294	0.52	0.16

Tableau 1. Résultats de clustering sur le corpus de données thématiquement homogènes et sur le corpus de données polythématiques pour l'ensemble des méthodes testées.

L'analyse des mesures de qualité sur le premier corpus de données homogènes met en évidence de bons résultats pour presque toutes les méthodes neuronales, excepté pour la méthode I^2GNG. Les meilleurs résultats sont obtenus avec les méthodes basées sur une topologie libre, et particulièrement avec la méthode NG et avec la méthode IGNG. De plus, étant donné que la puissance de la synthèse d'une méthode de clustering est aussi liée à sa capacité à produire de bons résultats avec un nombre de clusters aussi faible que possible, l'avantage tourne nettement en faveur d'IGNG qui obtient une Micro-F-mesure pratiquement équivalente à celle du NG, tout en nécessitant un nombre de clusters deux fois moindre. Un autre résultat intéressant est également mis en lumière par cette expérience : dès lors que l'on peut optimiser son nombre de clusters, comme c'est le cas dans le cadre de cette expérience, l'algorithme d'apprentissage de NG semble être plus efficace que celui de GNG (nombre final de clusters similaire pour les deux méthodes avec de meilleures valeurs de qualité pour la méthode NG). Dans le cas de la méthode non neuronale K-MEANS, il apparait une différence assez nette entre la Macro F-mesure et la Micro F-mesure. Comme nous l'avons décrit dans la section 3, cette différence illustre des résultats de clustering déséquilibrés comprenant des clusters bruités de taille importante[4]. Les résultats d'I^2GNG apparaissent seulement moyens, alors qu'ils devraient être au moins équivalents à ceux d'IGNG, étant donné que ce premier algorithme est censé représenter une amélioration du second. La méthode I^2GNG semble donc clairement souffrir de problèmes d'initialisation, comme cela pouvait être pressenti. En effet, comme aucune valeur par défaut n'est fournie par cette méthode pour définir l'influence de voisinage d'un cluster nouvellement créé (celui-ci ne contient alors qu'une seule donnée associée), chaque cluster peut initialement agglomérer aléatoirement n'importe quel type de données, y compris des données qui sont très dissimilaires les unes des autres. Ceci peut par conséquent mener à la création des clusters hétérogènes de taille relativement importante qui peuvent à leur tour coexister avec des clusters de très petite taille, notamment si des groupes de données suffisamment discriminants sont présents dans le corpus d'apprentissage.

Une première analyse des résultats sur notre deuxième corpus de données polythématiques prouve que la plupart des méthodes de clustering ont des difficultés énormes à traiter ce type de données, produisant par conséquent des résultats de qualité très médiocre, même pour leur nombre optimal de clusters. Ces mauvais résultats sont illustrés plus particulièrement par de faibles valeurs de Micro F-mesure. La différence très élevée entre les valeurs de Macro F-mesure et de celles de Micro F-mesure illustre quant à elle la présence de clusters-poubelles attirant la majeure partie des données, parallèlement à celle de clusters-poussières représentant des groupes marginaux, ou encore non complètement formés. C'est très nettement le cas pour les méthodes K-MEANS et I^2GNG, et à une plus faible ampleur pour la méthode NG. La méthode IGNG qui a produit les meilleurs résultats avec des données homogènes devient également fortement concernée par ce phénomène à partir du moment où elle doit gérer des données polythématiques. Le seul comportement relativement cohérent sur ces données est celui de la méthode SOM. Celui-ci est illustré par une valeur moins faible de Micro F-mesure et par une bonne concordance entre la valeur de Micro F-mesure et celle de Macro F-mesure. Un cas très critique est celui de méthode K-MEANS, dont le résultat revient en définitive à agglomérer l'ensemble des documents du corpus dans un seul cluster de taille significative[5]. Une situation également critique se produit avec la méthode GNG qui ne fournit aucun résultat sur ce corpus en raison de son incapacité à s'échapper d'un cycle infini de la création-destruction de neurones, autrement dit de clusters.

Les résultats ainsi que les comportements des méthodes peuvent être confirmés par une expertise approfondie du contenu des clusters et de la

[4] Ce phénomène va s'accentuer pour cette méthode et toucher quasiment toutes les autres méthodes dans le cas du traitement de données hétérogènes.

[5] Dans le cas du corpus polythématique, la méthode K-MEANS ne produit systématiquement qu'un seul cluster significatif associé à des clusters-miettes ne contenant qu'un seul document (un cluster de 1189 documents et 154 clusters d'1 seul document, dans notre cas). Ce qui revient à dire que dans ce cas, cette méthode à un pouvoir de discrimination nul.

distribution des tailles de ces derniers. Pour faciliter ce travail, il est également possible de juger de la nature des résultats en s'intéressant aux étiquettes qu'il est possible d'associer aux clusters en employant la stratégie de maximisation de vraisemblance décrite par l'équation 8. Les résultats des étiquetages obtenus par cette stratégie pour les cas respectifs des méthodes K-MEANS et SOM sont présentés aux figures 2 et 3.

[1189] 840-- Racine Densité Protéine Loi échelle Appareil respiratoire pathologie Condition aux limites Forêt Protéine lait Protection environnement Prévision Analyse donnée Pollution air Fibre In situ DNA Alcool Mécanique roche Sol Méthode analytique Cosmologie Traitement donnée Cartographie Végétation Rhizosphère Eau potable Application Méthode Monte Carlo Zone urbaine Grain Pédiatrie Cours eau Logiciel Partie aérienne végétal Inhibiteur enzyme Diagramme phase Photosynthèse Système nerveux pathologie Expression génique Variation saisonnière Nucléosynthèse Biodisponibilité Matière organique Ingénierie Vérification programme Cytométrie flux Lymphocyte T Diffusion(transport) Floculation Analyse composante principale Déprédateur Détection Produit contraste Arbre forestier feuillu Tolérance Industrie alimentaire Système à retard Tolérance faute Système temps réel Problème NP difficile Appareil respiratoire Polymère Synchronisation .....

Figure 2. Vue générale des étiquettes d'un cluster généré par la méthode K-MEANS. Ce cluster regroupe 1189 documents sur les 1341 du corpus, et 840 étiquettes sur les 889 de fréquence supérieure 3. Il attire un grand nombre d'étiquettes de différentes natures, figurant un contenu très hétérogène.

[28] 27-- Etude théorique Méthode dynamique moléculaire Simulation numérique Interface gaz liquide Continuum Modèle thermodynamique Fluide non newtonien Diffusion Nanoparticule Laser semiconducteur Propriété mécanique Analyse statistique Système binaire Compressibilité Point critique Propriété rhéologique Ecoulement conduite Perte charge Macromolécule Effet dimensionnel Mouillage

[21] 38-- Zone tempérée Foresterie Arbre forestier feuillu Arbre forestier Peuplement forestier Plante ligneuse Forêt décidue Montagne Sylviculture Futaie Arbre forestier résineux Formation végétale Gestion forestière Dendrométrie Peuplement forestier artificiel Filière bois Ecophysiologie Stade juvénile plante Entomologie Phytopathologie Caractéristique peuplement Industrie bois Qualité production Tolérance Economie forestière Dépérissement forêt Modèle simulation Peuplement forestier mélangé Symptomatologie Recommandation Bois feuillu Histoire Réglementation Bois résineux Contrainte Photosynthèse Climat Canopée

[16] 28-- Cardiopathie Insuffisance cardiaque Aigu Conduite à tenir Myocarde pathologie Pronostic Mortalité Chimiothérapie Epidémiologie Appareil circulatoire Chronique Evolution Cœur Stratégie Electrocardiographie Santé publique Electrodiagnostic Facteur prédictif Maladie Prévalence Soin Complication Coût Maladie héréditaire Tomodensitométrie Résultat Risque

Figure 3. Vue générale des étiquettes extraites de la partition générée par la méthode SOM, précédées du nombre de documents et du nombre d'étiquettes par classe. Il y a ici différents clusters de taille semblable attirant des groupes d'étiquettes sémantiquement homogènes.

Une cause plausible du problème d'agglomération de données, se produisant avec presque toutes les méthodes de clustering examinées avec ce corpus, est la nature de la similarité exploitée par défaut par l'ensemble des méthodes expérimentées pendant l'apprentissage, en l'occurrence la distance euclidienne. En effet, c'est un phénomène connu que cette distance, qui est la mesure de similarité la plus utilisée pour le clustering, devient faiblement discriminante dans les espaces fortement multidimensionnels contenant des données éparses [30] : dans ce contexte la distribution des distances euclidiennes est fortement biaisée vers les petites valeurs, ce qui revient également à considérer les données comme très similaires entre elles. Seul l'apprentissage sous contrainte de grille proposé par la méthode SOM de Kohonen semble représenter une bonne stratégie pour préserver celle-ci de produire des résultats trop mauvais dans un contexte aussi critique. De fait, ce type d'apprentissage impose l'homogénéité des résultats en répartissant à la fois les données et le bruit sur la grille.

Afin de résoudre les problèmes observés sur le corpus de données polythématiques, nous proposons à la section suivante plusieurs adaptations applicables sur les algorithmes de clustering potentiellement les meilleurs sur les données homogènes.

5. Adaptations originales des algorithmes IGNG et I2GNG

5.1 Adaptations d'IGNG

Choix aléatoire du gagnant final parmi des gagnants multiples (IGNG-R) : Du fait de l'utilisation d'une distance non discriminante, nous avons pu vérifier expérimentalement qu'à beaucoup d'étapes de l'apprentissage de l'algorithme IGNG il existait plusieurs neurones concurrents qui pouvaient être considérés comme des gagnants potentiels, étant donné que ces neurones avaient la même distance vis-à-vis de la donnée d'entrée courante. L'algorithme doit cependant choisir un gagnant final du fait de sa stratégie de type « le gagnant emporte le plus ». Dans l'algorithme original, chaque fois que cette situation se présente, la stratégie appliquée par défaut consiste à prendre comme gagnant final le premier neurone de la liste des neurones créés. Dans le cas du corpus de données polythématiques ce problème s'est révélé être très fréquent, et conduit directement à l'apparition d'un cluster-poubelle, représenté par le premier neurone créé. Une meilleure stratégie consiste à effectuer un processus de sélection aléatoire dans le cas où un gagnant doit être choisi parmi plusieurs neurones équidistants d'une donnée d'entrée. La conséquence indirecte de cette nouvelle stratégie est également celle de repartir le bruit entre plusieurs neurones.

Choix du gagnant final parmi des gagnants multiples employant une approche de maximisation de l'étiquetage des clusters (IGNG-M) : Chaque fois qu'un gagnant doit être choisi parmi plusieurs neurones équidistants d'une donnée d'entrée, une stratégie plus pertinente que l'approche aléatoire consiste à employer une approche de maximisation de la F-mesure de l'étiquetage pour choisir le gagnant final. Selon cette stratégie, la donnée d'entrée est d'abord attachée à tous les challengers potentiels pour créer des « challengers-augmentés », et la F-mesure moyenne d'étiquetage des « challengers-augmentés » est calculée sur la base de l'équation 8. Le gagnant est ensuite choisi comme le challenger dont l'état augmenté maximise positivement la différence de F-mesure moyenne d'étiquetage avec son état standard.

Utilisation directe d'une approche de maximisation de la qualité de l'étiquetage des clusters comme mesure de similarité (IGNG-F) : La dernière stratégie présentée peut malgré tout ne pas permettre de faire face à l'influence critique de la distance euclidienne dès lors que, dans cette stratégie, cette distance non discriminante reste employée comme critère de sélection principal. Une stratégie qui semble plus adéquate consiste

donc à supprimer complètement l'utilisation de la distance euclidienne dans le processus de sélection d'un gagnant, en considérant cette fois l'approche de maximisation de la qualité de l'étiquetage des clusters comme seul processus de sélection. Un autre avantage d'une telle stratégie est de fournir pour la première fois à la l'approche IGNG un caractère véritablement incrémental. En effet, ce nouvel algorithme fonctionne sans ne nécessiter aucun paramètre prédéfini dépendant du corpus analysé ou de données initiales. Un autre de ses avantages déterminants pour l'incrémentalité est qu'il peut prendre en compte un espace de description des données d'entrée ouvert, ou évolutif.

5.2 Adaptation d'I2GNG

Stratégie d'initialisation de réciprocité de voisinage (I²GNG-N) : Une approche possible pour faire face aux problèmes d'initialisation de la méthode I²GNG peut consister à créer des noyaux d'initialisation afin de fournir des paramètres initiaux d'influence de voisinage aux clusters. Dans ce contexte, une stratégie d'initialisation basée sur les voisins réciproques [4] semble être l'une des plus appropriées. En effet, cette stratégie présente l'avantage de minimiser l'influence de l'initialisation sur l'évolution ultérieure des clusters. Néanmoins, un des inconvénients principaux d'employer une stratégie d'initialisation, quelle qu'elle soit, est de supprimer le caractère incrémental de cet algorithme.

5.3 Nouveaux résultats

Le tableau 2 illustre les résultats obtenus avec nos nouvelles stratégies. Il démontre plus particulièrement l'efficacité de notre stratégie de clustering basée directement sur la maximisation de l'étiquetage (IGNG-F) qui surpasse toutes les stratégies existantes, en produisant les résultats les meilleurs sur les données polythématiques, et notamment avec le plus petit nombre de clusters. On peut également remarquer qu'une amélioration significative des performances de l'algorithme I²GNG est obtenue une fois qu'une stratégie d'initialisation du voisinage des clusters est employée (I²GNG-N).

METHODE DE CLUSTERING	NBR OPTIMAL DE CLUSTERS	F-MESURE MICRO
SOM (Référence)	**289**	**0.40**
IGNG Original	378	0.21
IGNG-R (Affectation aléatoire des égaux)	382	0.43
IGNG-L (Affectation des égaux par maximisation d'étiquetage)	437	0.44
IGNG-F (Similarité basée sur la maximisation d'étiquetage)	**198**	**0.47**
I²GNG Original	294	0.15
I²GNG-N (Initialisation par voisinage réciproque)	221	0.38

Tableau 2. Résultats de clustering sur le corpus de données polythématiques pour les nouveaux algorithmes développés.

6. Conclusion

Dans ce travail, nous avons montré que les méthodes de clustering neuronales, comme la méthode IGNG, affichaient de très bonnes performances dans le contexte de l'analyse d'un corpus de données textuelles homogènes. Pour évaluer les performances de ces méthodes nous avons exploité des mesures spécifiques de qualité qui présentent la propriété d'être indépendantes de la méthode considérée. En exploitant ces mêmes mesures, nous avons également montré la diminution drastique de performance de la plupart des méthodes de clustering, quand un corpus de données textuelles polythématiques est considéré en entrée.

Ce problème peut s'avérer critique dans le cadre de nombreuses applications, comme les applications de veille, et en particulier celles de plus en plus nombreuses, qui sont amenées à manipuler des données polythématiques, hétérogènes ou changeantes. Il y avait donc nécessité de réviser en profondeur le mode opératoire des méthodes de clustering. En nous basant sur certains des principes de nos mesures de qualité, nous avons proposé un algorithme de gaz neuronal croissant original exploitant une nouvelle mesure de similarité basé sur la maximisation de la qualité de l'étiquetage des clusters, comme alternative à la distance euclidienne exploitée dans les approches usuelles. Nous avons expérimentalement prouvé que cette nouvelle méthode fournissait des résultats supérieurs aux méthodes de référence existantes en contexte polythématique. Par ailleurs, il apparait qu'elle fournit également un véritable caractère incrémental à l'algorithme que nous avons proposé.

Ces dernières possibilités nous ouvrent des perspectives claires pour appliquer dans un futur proche notre nouvelle approche de clustering incrémentale au domaine de l'exploitation des données textuelles changeant au cours du temps. Dans un futur plus lointain, nous comptons également appliquer cette approche, en l'associant au paradigme d'analyse de données multi-vues que nous avons précédemment proposé [20], à d'autres domaines porteurs mettant en jeux des données numériques en parallèle avec des données textuelles, tout en intégrant une dimension temporelle, comme c'est le cas des données génomiques.

Nous comptons également adapter le principe original de similarité par maximisation d'étiquetage des clusters à plusieurs autres algorithmes de clustering. La latitude d'évolution de notre nouvelle approche pour l'analyse prospective et pour la veille est donc particulièrement grande.

7. Bibliographie

[1] Attik M., Al Shehabi S. and Lamirel J.-C. (2006). Clustering Quality Measures for Data Samples with Multiple Labels. Proceedings of the The IASTED International Conference on Databases and Applications (DBA), Innsbruck, Austria, February 2006.

[2] Davies, D.and Bouldin, W. (2000). A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell, 1(4), 224-22.

[3] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood for incomplete data via the em algorithm. Journal of the Royal Statistical Society, vol. B-39: 1-38.

[4] De Rham C. (1980). La classification hiérarchique ascendante selon la méthode des voisins réciproques. Les cahiers de l'analyse de données, Vol. 5, No. 2, pages 135-144, 1980.

[5] Dunn J. (1974): Well Separated clusters and optimal fuzzy partitions. Journal of Cybernetics,4, 95-104.

[6] Fritzke B. (1995). A growing neural gas network lerans topologies. Tesauro G., Touretzky D. S., leen T. K., Eds., Advances in neural Information processing Systems 7, pp 625-632, MIT Press, Cambridge MA.

[7] François C., Hoffmann M., Lamirel J.-C. and Polanco X. (2003). Artificial Neural Network mapping experiments. EICSTES (IST-1999-20350) Final Report (WP 9.4), 86 p., September 2003.

[8] Ghribi M., Cuxac P., Lamirel J.C., Lelu A. (2010). Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés. Atelier EGC'EVALECD, Hammamet, Tunisia, January 2010.

[9] Hamza H., Belaïd Y., Belaïd. A and Chaudhuri B. B. (2008). Incremental classification of invoice documents. 19th International Conference on Pattern Recognition - ICPR 2008.

[10] Hebb, D.O. (1949). The Organization of Behavior: A Neuropsychological Theory, Wiley, New York, 1949.

[11] Jouve, O. (1999). Les nouvelles technologies de la recherche d'information, Séminaire Documentation, Paris.

[12] Kaski S., Honkela T., Lagus K. and Kohonen, T. (1998). WEBSOM-self organizing maps of document collections, Neurocomputing, vol. 21, pp. 101-117.

[13] Kassab R. and Lamirel J.-C. (2008). Feature Based Cluster Validation for High Dimensional Data. IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008.

[14] Kohonen T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, vol. 43, pp 56-59.

[15] Kohonen T. (2001). Self-Organising Maps. 3rd ed. Berlin: Springer-Verlag, 2001.

[16] Lamirel J.-C. and Créhange M. (1994). Application of a symbolico-connectionist approach for the design of a highly interactive documentary database interrogation system with on-line learning capabilities. Proceedings ACM-CIKM 94, Gaitherburg, Maryland, USA, November 94.

[17] Lamirel, J-C., Ducloy J. and Oster, G. (2000). Adaptative browsing for information discovery in an iconographic context, In Conference Proceedings RIAO, Paris, Vol. 2, p. 1657-1672.

[18] Lamirel J.-C. and Al Shehabi S. (2003). Neural Network driven Patent Analysis. Proceedings of ACL 2003 Workshop on Patent Analysis, Sapporo, Japan, July 2003.

[19] Lamirel J.-C., Al-Shehabi S., François C. and Hoffmann M. (2004a). New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. Scientometrics, 60(3).

[20] Lamirel J.-C., Al Shehabi, S., François, C., & Polanco, X. (2004b). Using a compound approach based on elaborated neural network for Webometrics: an example issued from the EICSTES Project. Scientometrics, Vol. 61, No. 3, p. 427-441.

[21] Lamirel J.-C., Ta A.P. and Attik M. (2008). Novel Labeling Strategies for Hierarchical Representation of Multidimensional Data Analysis Results. IASTED International Conference on Artificial Intelligence and Applications (AIA), Innsbruck, Austria, February 2008.

[22] Lebart L.,Maurineau A., Piron M. (1982): Traitement des données statistiques. Dunod, Paris.

[23] Martinetz T. and Schulten K. (1991). A "neural gas" network learns topologies. In Kohonen, T., Makisara K., Simula O., and Kangas J., editors, Articial Neural Networks, pp 397-402. Elsevier Amsterdam.

[24] Merkl D., Shao H.H., Dittenbach M. and Rauber A. (2003). Adaptive hierarchical incremental grid growing: an architecture for high-dimensional data visualization. In Proceedings of the 4th Workshop on Self-Organizing Maps, Advances in Self-Organizing Maps, pp 293-298, Kitakyushu, Japan, September 11-14 2003.

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

[26] Prudent Y. and Ennaji A. (2005). An Incremental Growing Neural Gas learns Topology. ESANN2005, 13th European Symposium on Artificial Neural Networks, Bruges, Belgium, 27-29April 205, published in Neural Networks, 2005. IJCNN apos;05. Proceedings. 2005 IEEE International Joint Conference , vol. 2, no. 31 pp 1211 - 1216, July-4 Aug. 2005.

[27] Robertson S. E. and Sparck Jones, K. (1976). Relevance Weighting of Search Terms. Journal of the American Society for Information Science, 27:129–146.

[28] Rousseeuw P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[29] Varsis A. and Versino C. (1992). Clustering of Socio-Economic Data with Kohonen Maps, In Proceedings of third International Workshop on Parallel Applications in Statistics and Economics, Pragues, Czechoslovakia.

[30] Verleysen, M. (2004). Learning High-Dimensional Data with Artificial Neural Networks, LEARNING'04, Elche, Spain, 20-22 October 2004.