Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés

Maha Ghribi, Pascal Cuxac, Jean-Charles Lamirel, Alain Lelu

Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés

M. GHRIBI[1], P. CUXAC[1], J.C. LAMIREL[2], A. LELU[2] [3]

[1]INIST-CNRS, 2 allée du Parc de Brabois, 54500-Vandœuvre-lès-Nancy, France.
maha.ghribi@inist.fr ; pascal.cuxac@inist.fr
[2]LORIA Campus Scientifique BP 239 - 54506 Vandœuvre-lès-Nancy, France.
alain.lelu@loria.fr ; jean-charles.lamirel@loria.fr
[3]LASELDI / Université de Franche-Comté, 30 rue Mégevand – 25030 Besançon, France.

Résumé : Nos travaux sur une nouvelle méthode de classification non supervisée (Germen) nous ont amenés à nous interroger sur la qualité des résultats obtenus. Le problème est d'estimer si une méthode de clustering est 'meilleure' qu'une autre pour le type de données que nous traitons (données textuelles). Dans un premier temps, après avoir fait un état de l'art des méthodes existantes, nous avons appliqué quelques indices de qualité aux résultats de clustering issus de notre algorithme Germen ainsi que d'autres algorithmes communément utilisés. Ces indices de qualité ne permettant pas de sélectionner la meilleure partition, nous avons développé une nouvelle série d'indices basés sur la distribution des mots-clés. Nous présentons et discutons les résultats obtenus ainsi que les réflexions engagées pour faire évoluer l'évaluation de classifications non supervisées sur des textes.

1 Introduction

Evaluer les performances d'un algorithme de clustering (classification non supervisée) n'est pas chose aisée. Une première façon est de faire une évaluation supervisée : on compare le résultat obtenu à une référence [Yosr et Sinaoui 2009] (on peut pour cela utiliser une classification préexistante ou des corpus de référence). En ce qui concerne les corpus de tests exploitables dans le cadre de l'évaluation des méthodes de classifications, outre les corpus de test numériques classiques utilisés pour les méthodes supervisées tels qu''Iris' ou 'Mushroom', des corpus de textes indexés issus de dépêches 'Reuters' sont régulièrement utilisés dans différentes campagnes d'évaluation. Dans ce dernier cas, notre expérience montre cependant que l'indexation proposée par [Lewis et al 2004] est peu adaptée à l'évaluation de méthodes de classifications non supervisées de données textuelles [Cuxac et al. 2009].

Nous proposons de mesurer la qualité de méthodes de clustering que nous avons développées en les comparant à des méthodes choisies dans la littérature. Pour cela nous avons sélectionné quelques indicateurs de qualité communément utilisés ; leur application à nos résultats nous a amenés à une réflexion sur ces indicateurs et à faire de nouvelles propositions.

2 Les données

Nous avons utilisé un corpus bibliographique test issu de la base PASCAL, édité par L'INIST (CNRS, www.inist.fr). Ce corpus rassemble 1920 notices sur le thème de « la recherche en Lorraine », signalant des documents de type article, congrès ou thèse. Pour réaliser les classifications, nous utilisons les termes d'indexation manuelle présents dans les notices. Le corpus est indexé par 3557 descripteurs de fréquence supérieure à 1. La figure 1 illustre quelques champs d'une notice bibliographique de ce corpus.

ET : Seismic and geotechnical investigations following a rockburst in a complex French mining district
AU : DRIAD-LEBEAU (L.); LAHAIE (F.); AL HEIB (M.); JOSIEN (J. P.); BIGARRE (P.); NOIREL (J. F.); HOWER (James C.); GREB (Stephen F.)
AF : Institut National de l'Environnement Industriel et des Risques (INERIS)-Ecole des Mines, Parc Saurupt/54042 Nancy/France (1 aut., 2 aut., 3 aut., 5 aut.); GEODERIS/Metz/France (4 aut.); Charbonnages de France/Merlebach/France (6 aut.); University of Kentucky Center for Applied Energy Research/Lexington, KY/Etats-Unis (1 aut.); Kentucky Geological Survey, University of Kentucky/Lexington, KY/Etats-Unis (2 aut.)
SO : International journal of coal geology; ISSN 0166-5162; Pays-Bas; Da. 2005; Vol. 64; No. 1-2; Pp. 66-78.
EA : This paper presents the results of seismic and geotechnical studies carried out after a fatal accident that occurred during mining of the Frieda5 coal seam at Merlebach mine of HBL (Houillères du Bassin Lorrain, East France). On June 21, 2001, a violent rockburst (local magnitude of 3.6) affected the Frieda5 seam at depth of approximately 1250 m …
ED : mining; coal seams; coal mines; stress; joints; sandstone; channels; rock bursts; seismic methods; safety; risk assessment; acoustical methods; France; Lorraine; Moselle France
EG : clastic rocks; sedimentary rocks; Western Europe; Europe

FIG. 1- *Une notice bibliographique de la base PASCAL (ET = titre ; AU = auteurs ; AF = affiliations ; SO = source ; EA= résumé ; ED+EG = descripteurs)*

Un tel corpus est ensuite transformé en une matrice booléenne documents×mots, très creuse par nature. Bien que ces données ne proviennent pas directement de textes, nous avons constaté que la répartition des termes avait l'allure zipfienne (relation linéaire entre le log des fréquences de termes et le log de leur rang) caractéristique des données textuelles. Ce corpus peut-être obtenu en contactant les auteurs.

3 Les méthodes de clustering et leurs résultats

3.1 Méthodes de clustering utilisées

Dans cette partie, nous allons décrire les méthodes de clustering utilisées : des méthodes neuronales (SOM, IGNG, K-means Axiales, Analyse en Composantes locales++), et des méthodes opérant sur des graphes (Walktrap, Germen).

3.1.1 Méthodes neuronales

Nous nous sommes intéressés à quatre méthodes qui sont les K-means Axiales (KMA), l'Analyse en Composantes Locales++ (ACL++), SOM (carte auto-organisatrice de Kohonen) et IGNG (Incremental Growing Neural Gas)

Dans **la méthode KMA** (K-Means Axiales) [Lelu 1994], les vecteurs-documents sont normalisés selon la métrique de Hellinger, particulièrement adaptée aux données textuelles, et sont affectés à des axes de classe (« vecteurs-neurones ») pointant vers les zones de forte

densité des données, avec des degrés de centralité dans leur classe plus ou moins prononcés selon le principe des centres mobiles des K-means.

La méthode ACL++ (Analyse en composantes locales dans l'espace sphéré des données) : L'Analyse en composantes locales (Lelu 1994) est comme Germen, décrite plus loin, une méthode de détermination d'axes de clusters par montée en gradient de vecteurs-neurones, sur le paysage de densité des données. La densité est déterminée en tout point de cet espace à partir d'un voisinage de rayon fixe, à la différence de Germen où la densité est adaptative.

La variante ACL++, à la différence de l'ACL, n'est pas réalisée dans l'espace des descripteurs, mais dans l'espace des K premiers vecteurs singuliers de la matrice des données, où K est déterminé par un test statistique [Lelu et Cadot,2010]. Elle aboutit à une "normalisation" des densités dans cet espace et à une adaptativité aux différences de densité, de façon moins locale que dans Germen.

La méthode SOM [Kohonen,1982] est une méthode où les vecteurs-neurones s'auto-organisent au fil des données sous forme d'une structure de voisinage bien définie (« ficelle » unidimensionnelle, grille bidimensionnelle ou multidimensionnelle). Considérée comme une méthode statique, c'est-à-dire à nombre de neurones pré-fixé, son algorithme débute par l'initialisation de la carte de voisinage avec une sélection aléatoire des neurones. A chaque itération, l'insertion d'une nouvelle donnée induit un auto-ajustement de la carte par la modification du vecteur de référence du neurone le plus proche de la donnée d'entrée ainsi que ses voisins directs.

La méthode IGNG [Prudent, Ennaji,2004], [Prudent, Ennaji,2005a], [Prudent, Ennaji,2005b] est une méthode neuronale très différente de SOM puisque d'une part, elle n'impose aucune structure de voisinage et d'autre part, le nombre de neurones varie au cours de l'apprentissage. Il y a possibilité de suppression et de création de neurones. La suppression est liée à l'introduction de la notion d'âge aux liens entre les neurones. Si à une itération l'âge d'un lien atteint la maturité, celui-ci est supprimé. Les neurones se trouvant isolés sont automatiquement supprimés. La création se fait à chaque itération si certaines conditions sont vérifiées. En effet, avec l'introduction d'une nouvelle donnée, et avant d'effectuer l'apprentissage, on vérifie si la création d'une nouvelle classe est nécessaire. La règle est la suivante : s'il existe au moins deux neurones dont la distance par rapport à la nouvelle donnée est inférieure à une certaine valeur σ préfixée, la création du neurone est inutile. Sinon, la nouvelle donnée représente le nouveau neurone. σ est calculé à partir de toutes les données présentes dans l'échantillon. C'est la distance moyenne qui sépare toutes les données de la donnée centrale.

Une caractéristique très intéressante de IGNG, en dehors de son aspect dynamique, est sa tolérance aux données bruitées. En effet, l'algorithme affecte à chaque neurone un âge depuis sa création. Si à une certaine itération, un neurone atteint l'âge de maturité, il passe de l'état d'un neurone embryon à un neurone mature. L'âge de maturité correspond donc au nombre minimal d'activation pour qu'un neurone ne soit plus considéré comme résultant de données bruitées.

3.1.2 Méthodes opérant sur des graphes

Parmi les méthodes qui opèrent sur des graphes, nous nous somme intéressés à deux méthodes, Walktrap et Germen.

Walktrap [Pons, Latapy : 2006] ou méthode des marches aléatoires (Random Walks) a pour but de décomposer le graphe en un certain nombre de « communautés » ou classes. Son principe est similaire à celui de la classification ascendante hiérarchique. En effet, commençant par une partition où chaque donnée forme une classe, à chaque paquet de t itérations de marches aléatoires, elle fait fusionner les deux classes qui d'une part, présentent au moins un lien entre leurs données et d'autre part, en se basant sur la méthode de Ward, minimisent la moyenne de la distance au carré de chaque sommet à sa communauté. La différence de Walktrap par rapport à la méthode ascendante hiérarchique est qu'elle calcule la distance entre les nœuds du graphe à partir de la matrice d'adjacence. Cette dernière permet de déterminer la matrice de transition entre les éléments du graphe (nœuds et communautés). L'algorithme se termine par une partition qui contient une seule classe regroupant tous les nœuds. Le choix de la meilleure partition est fait de façon à ce quelle maximise le critère de « Modularité » [Newman, Girman, 2004] décrit plus bas.

Germen [Cuxac, et al 2009], [Lelu et al. 2006] est une méthode de clustering de graphes. Elle se base sur la notion de densité du nuage des vecteurs-données. Son principe se résume en la détection des maxima de densité à chaque itération et en la prise en compte des perturbations locales dues à l'introduction d'un nouvel élément. A chaque élément du graphe (document) on attribue une densité. Le partitionnement du graphe se fait par le repérage des nœuds les plus denses. Ces derniers représentent les « chefs de classes ». Leurs zones d'influence s'élargissent au fur et à mesure par rattachement unique ou partagé de leurs voisins de plus en plus éloignés. Ce rattachement se base sur la notion d'héritage des étiquettes des chefs de classes. Plusieurs règles peuvent exister, par exemple :

- Un nœud hérite du numéro du chef de classe de son voisin le plus surplombant (ayant une densité la plus importante que la sienne et son 1-voisinage). Si celui-ci n'existe pas, on crée une nouvelle classe. Dans ce cas, un document ne peut appartenir qu'à une seule classe (classification non recouvrante).
- Un nœud hérite des numéros de chefs de classes de tous ses voisins surplombants. Dans ce cas, un document peut appartenir à plusieurs classes (classification recouvrante).

Pour la construction du graphe, Germen peut utiliser différentes méthodes (K plus proches voisins par exemple) mais, suite à nos expériences précédentes [Cuxac et al. 2006], nous leur avons préféré ici la méthode « Tournebool » [Cadot, 2006]. C'est un algorithme de validation statistique des liens entre vecteurs-données qui consiste en la génération d'un grand nombre de matrices booléennes de sommes marginales équivalentes à celles de la matrice des données. Pour chaque couple de documents on calcule son support (nombre de mots clés commun) dans chaque matrice générée. On obtient ainsi une distribution des supports des deux documents. Si le support dans la matrice de données initiale est supérieur à un certain seuil prédéfini dans la distribution, le lien est considéré non dû au hasard, donc valide.

3.2 Résultats de clustering

Pour toutes les méthodes on a fait varier les paramètres pour avoir plusieurs partitions avec des nombres de classes différents, par exemple le nombre de classes pour les méthodes qui le considèrent comme paramètre, comme SOM et KMA. Alors qu'ACL++ fixe le nombre de classes à la seule « dimension structurelle » déterminée par son test statistique, ici autour de 155, on a fait varier le sigma d'IGNG. Avec Germen, on a seuillé les liens, ce qui a permis d'obtenir des partitions différentes. Avec Walktrap, on a varié le nombre d'itérations.

Méthode	Nombre de Classe	Nbr de docs classés	Taille de la plus grosse classe	Nbr de classes de taille < 10	Nbr de classes de taille > 100
SOM	49	1920	138	21	3
Germen	61	1707	450	46	4
Walktrap	67	1624	309	46	5
KMA	155	1920	108	77	1
ACL++	151	1920	52	78	0
IGNG	169	1920	338	110	1

TAB. 1 – *Résultats de classification des différentes méthodes de clustering*

Le tableau 1 montre pour chaque méthode la partition qui peut être considérée comme la meilleure en se basant sur les indices de qualité présentés par la suite. Pour les méthodes Germen et Walktrap, le nombre de documents classés est inférieur à la taille de l'échantillon utilisé : la construction du graphe a isolé certains documents qui ne sont pas considérés dans la classification. On remarque un comportement équivalent entre ces deux méthodes en termes de nombre de classes et taille des classes. Avec la présence de très grosses classes, on attend une hétérogénéité à l'intérieur de ces classes. Les méthodes KMA, ACL++ et IGNG présentent un nombre très important de classes de faible taille. Ceci risque d'influencer la répartition des similarités entre les classes. SOM avec une partition de 49 classes présente un équilibre de point de vue structure des classes (absence de très grosses classes et faible proportion de classes de petite taille). ACL++ également.

4 Les indices de qualité : un état de l'art

On distinguera dans cette brève revue deux familles d'indices de qualité de clustering : les indices opérant sur des graphes et ceux basés sur les distances.

4.1 Indices de qualité opérant sur les graphes

Les indices de qualité opérant sur des graphes s'intéressent aux liens entre les nœuds à l'intérieur des graphes. Ils se basent sur le principe que les nœuds appartenant à une même classe sont plus liés entre eux qu'avec les points appartenant à des classes différentes. Plusieurs formalisations ont été développées : la « Performance » et la « Modularité » utilisent des critères inter et intra classes afin de mesurer à quel point les classes sont formées par des éléments homogènes, et les classes bien séparées.

Dans cette partie, on note G(E,V) un graphe où E représente la liste des *n* nœuds et V représente la liste des arêtes dans le graphe. On note aussi $P=(C_1,\ldots,C_k)$ une partition du graphe.

La « Performance » [Van Dongen, 2000] introduit la notion de couple de nœuds correctement interprétés qui désigne à la fois les couples de nœuds liés appartenant à une même classe et les couples de nœuds non liés appartenant à deux classes différentes. La Performance calcule la fraction des couples de nœuds correctement interprétés par rapport au nombre total de couples de nœuds : $Performance(P) = \frac{m(P) + \sum_{\{u,v\} \notin E, u \in C_i, v \in C_j, i \neq j} 1}{\frac{1}{2} n(n-1)}$

$m(P)$ représente le nombre de liens à l'intérieur des classes.

Une augmentation de la Performance signifie que d'une part, les classes sont bien séparées et d'autre part, que les nœuds à l'intérieur des classes sont bien liés entre eux. Par conséquent, plus la performance est proche de 1 meilleure est la classification.

La « Modularité » [Newman, Girman ,2004] calcule, à coté de la proportion des liens intra classes, la proportion des liens inter classes.

$$Q(P) = \sum_{C \in P} (e_C - a_C^2)$$

e_C représente la fraction des liens Intra classe : $e_C = \frac{\sum_{u,v \in C, \{u,v\} \in E} 1}{|E|}$

a_C représente la fraction des liens qui ont au moins une extrémité dans C

$$a_C = \frac{\sum_{u \in C, v \in E, \{u,v\} \in V} 1}{|E|}$$

a_C fait intervenir les relations que possède la classe C avec les autres classes (les liens inter-classes).

Plus la proportion des liens intra-classe augmentent et les liens inter-classes diminuent, plus les documents à l'intérieur des classes sont liés et les classes sont séparées entre elles. Donc la meilleure partition c'est celle qui maximise la modularité.

4.2 Indices de qualité basés sur la distance

Les indices inertiels [Lebart et al, 1982] sont les plus connus et les plus utilisés pour évaluer la qualité d'une classification.

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

$$Intra = \frac{1}{n} \sum_{C \in P} \frac{1}{2n_c} \sum_{i \in C} \sum_{j \in C} d(i, j)^2$$

- 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.

$$Inter = \frac{1}{n} \sum_{C \in P} n_C d^2(c, c_G)$$

Avec c le centre de la classe C et c_G est le centre du nuage de points.

Plus les données à l'intérieur des classes sont homogènes, plus leurs distances par rapport au point représentant la classe sont faibles. Par conséquent, une valeur faible de l'inertie intra-classes décrit une homogénéité des données à l'intérieur des classes.

Plus les classes sont hétérogènes entre elles, plus les distances entres les points représentants les profils des classes sont élevées. Donc, une valeur élevée de l'inertie inter-classes traduit une hétérogénéité entre les classes. Cet indice a le défaut d'augmenter quand on augmente le nombre de classes.

Plusieurs autres indices de qualité qui utilisent la distance entre les individus ont été développés dont l'indice de Dunn, l'indice de validation de Davies–Bouldin et la Silhouette.

Les indices de Dunn et de Davies-Bouldin mélangent à la fois les inerties Intra classes et les inerties Inter classes.

L'indice de Dunn [Dunn,1974] est décrit par la formule suivante :

$$Dunn = \min_{1\le i\le n}\left\{\min_{\substack{1\le j\le n\\ j\neq i}}\left\{\frac{d(c_i,c_j)}{\max_{1\le k\le n}(d'(c_k))}\right\}\right\}$$

Il cherche la distance minimale qui sépare deux classes dans la partition tout en tenant compte de la distribution des éléments à l'intérieur des classes. Plus cette distance est grande meilleure est la partition.

L'indice de Davies-Bouldin (DB) [Davies et Bouldin , 2000] traite chaque classe individuellement et cherche à mesurer à quel point elle est similaire à la classe qui lui est la plus proche. L'indice DB est formulé de la façon suivante :

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left\{\frac{I(c_i)+I(c_j)}{I(c_i,c_j)}\right\}$$

Pour chaque classe i de la partition, on cherche la classe j qui maximise l' « indice de similarité » décrit comme suit $R_{ij} = \frac{I(c_i)+I(c_j)}{I(c_i,c_j)}$

$I(c_i)$ représente la moyenne des distances entre les documents appartenant à la classe C_i et son centre. Et $I(c_i,c_j)$ représente la distance entre les centres des deux classes C_i et C_j.

La meilleure partition est donc celle qui minimise la moyenne de la valeur calculée pour chaque classe. En d'autres termes, la meilleure partition est celle qui minimise la similarité entre les classes.

L'indice Silhouette [Rousseeuw, 1987] est différent, des indices de qualité traités ci-dessus ; il travaille à l'échelle microscopique, c'est à dire qu'il s'intéresse aux documents en particulier et non pas aux classes. Le but de Silhouette est de vérifier si chaque document a été bien classé. Pour cela, et pour chaque document i de la partition, on calcule la valeur suivante :

$$-1 \le S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \le 1$$

a(i) représente la distance moyenne qui le sépare des autres documents de la classe à laquelle il appartient et *b(i)* représente la distance moyenne qui le sépare des documents appartenant à la classe la plus proche.

Quand *S(i)* est proche de 1, le document est bien classé : la distance qui le sépare de la classe la plus proche est très supérieure à celle qui le sépare de sa classe. Par contre, si S(i) est proche de -1, cela veut dire que le document est mal classé. Mais si S(i) est proche de 0 alors il pourrait également être classé dans la classe la plus proche.

L'indice Silhouette de la partition est calculé à partir de la moyenne entre les indices de ses éléments.

Les mesures de qualité exposées ci-dessus sont intéressantes en soi, mais le domaine du traitement des données textuelles est un peu spécifique puisque les données ne sont pas décrites par des variables quantitatives mais plutôt par les descripteurs qualitatifs que sont les mots-clés caractérisant chaque document. Il n'est donc pas immédiat de définir les liens entre les documents afin d'utiliser les mesures de qualité qui se basent sur des graphes. De plus, le calcul des distances utilisables par les indices inertiels, les indices de Dunn, de Davies–

Bouldin et Silhouette sont délicats. En effet, les profils des documents sont des vecteurs binaires très creux dans un espace des mots-clés généralement de très grande dimension.

4.3 Les résultats

Nous allons maintenant appliquer les indices de qualité décrits précédemment sur les nos résultats de clustering. Pour cela, nous avons utilisé la distance de Jaccard pour calculer la distance entre les documents [Jaccard 1901]. Cette distance est adaptée aux données de profils binaires. Son principe est le suivant : pour chaque couple de documents D_i et D_j on a le tableau suivant :

$D_i \setminus D_j$	1	0
1	a	b
0	c	d

TAB. 2 – *Tableau croisé entre deux documents permettant de calculer une vaste famille d'indices de (dis-)similarité, en particulier leur distance de Jaccard*

a représente le nombre de descripteurs commun entre les deux documents, b représente le nombre de descripteurs présents dans *Di* et non pas dans D_i, c représente le nombre de descripteurs présents dans D_i et non pas dans D_i et d représente le nombre de descripteurs qui sont absents dans les deux documents.

La distance de Jaccard prend la forme suivante : $d(Di,Dj) = \frac{b+c}{a+b+c}$

Plus le nombre de descripteurs communs entre les documents est élevé plus la distance est faible.

Pour les méthodes de clustering opérant sur des graphes (Walktrap et Germen), on a construit le graphe avec la méthode de Tournebool décrite en 3.1.2 [Cadot, 2006].

Examinons maintenant le comportement des indices de qualité sur nos données documentaires. La Performance et l'inertie intra-classes illustrées dans la figure 2 montrent une forte dépendance avec le nombre de classes dans la partition. Selon ces deux indices, toutes les méthodes sont équivalentes et le choix de la meilleure partition devient difficile.

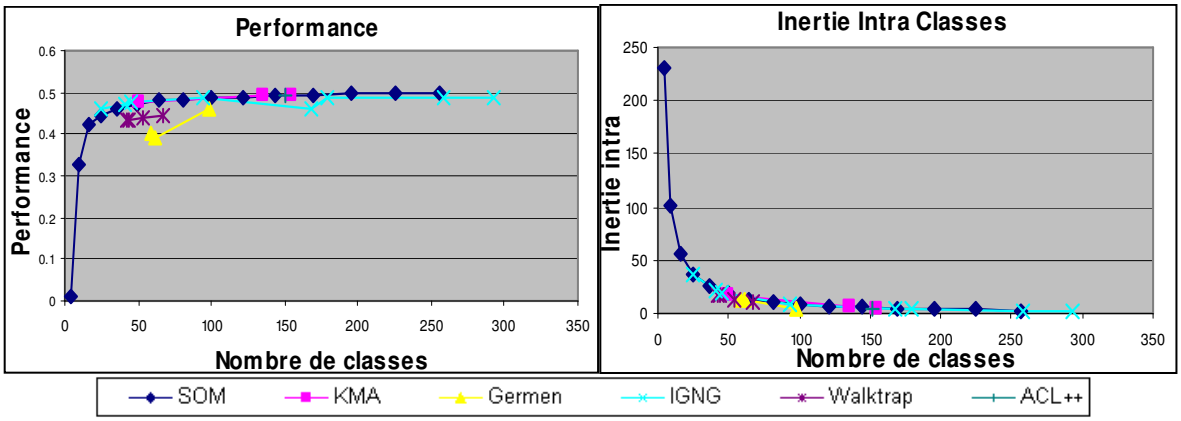

FIG. 2 – *Les valeurs de l'indice de la Performance et de l'Inertie Intra-Classes en fonction du nombre de classes produites par les différentes méthodes de clustering*

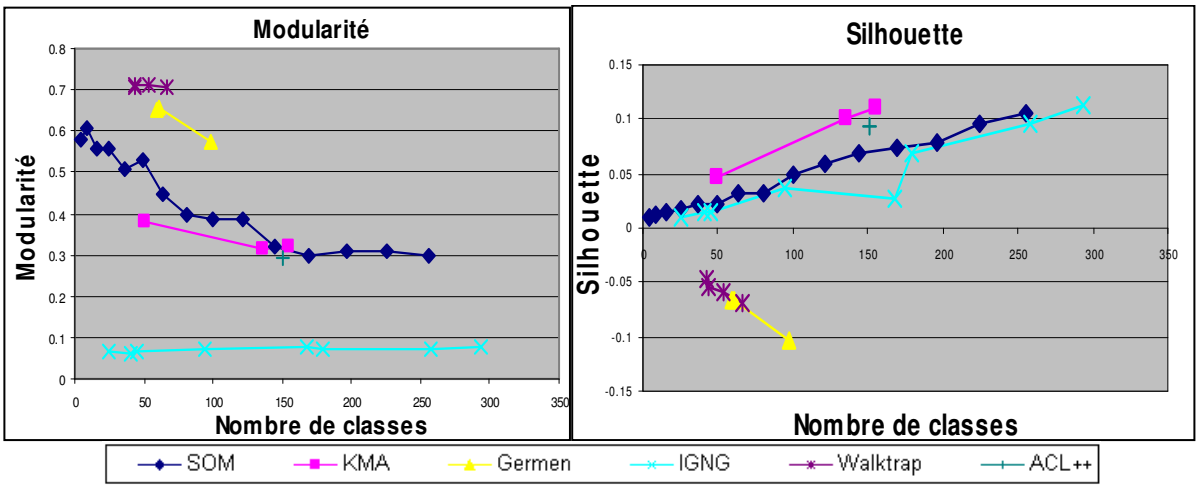

FIG. 3 – *Les valeurs de l'indice de Modularité et de Silhouette en fonction du nombre de classes produites par les différentes méthodes de clustering*

La modularité présente des valeurs très différentes selon les méthodes utilisées (figure 3). Cependant, on remarque que les valeurs associées à Germen et Walktrap sont les plus élevées. Ceci est attendu car elles présentent de très grosses classes et donc le nombre de liens intra classes est très important. Mais les résultats de la Modularité ne sont pas satisfaisants puisqu'ils n'ont pas permis de détecter l'hétérogénéité à l'intérieur des classes produites par ces deux méthodes.

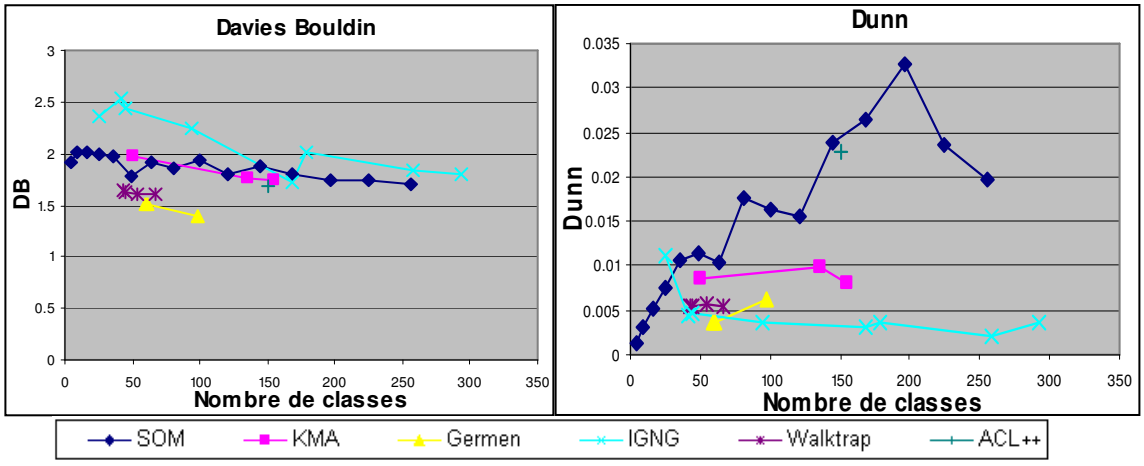

FIG. 4– *Les valeurs de l'indice de Davies Bouldin et de Dunn en fonction du nombre de classes produites par les différentes méthodes de clustering*

L'indice Silhouette illustré dans la figure 3 montre des valeurs absolues très proches de zéro. On peut conclure que les classes sont très proches les unes des autres et que les documents présentent une confusion au niveau de leur classification. Ce que l'on peut reprocher à Silhouette est que les documents appartenant à des classes caractérisées par une

grande dispersion de leurs éléments vont avoir des valeurs négatives s'ils sont proches de classes de faibles tailles. Ceci est clairement visible avec les partitions de Germen et Walktrap qui présentent à la fois des grosses classes et des classes de faible taille.

L'indice de Davies-Bouldin illustré dans la figure 4 montre une équivalence entre les résultats des deux familles de méthodes de clustering. Les valeurs sont très proches. Les méthodes neuronales présentent un comportement similaire de même que les méthodes Walktrap et Germen. L'indice de Dunn présente de très faibles variations avec des valeurs proches de zéro (figure 4).

5 De nouveaux indices de qualité : mesures basées sur les distributions des descripteurs

5.1 Rappel, précision, F-mesure

Les mesures de qualité habituellement utilisées ne sont donc pas optimales pour évaluer les résultats de classification non supervisée sur des corpus de textes. Notre démarche a alors consisté à développer des indices qui tiendraient mieux compte du type de nos données et qui seraient indépendants de la méthode de clustering utilisée, tout cela sans classification de référence.

La notion de Rappel, Précision et F-mesure a été introduite par Van Rijsbergen [Van Rijsbergen, 1979]. Elle se base sur le fait qu'un système de recherche documentaire est efficace s'il permet de restituer un maximum d'informations pertinentes. La Précision (P) détermine le pourcentage de documents pertinents restitués pour une requête donnée et le Rappel (R) calcule le pourcentage de documents pertinents restitués par rapport au nombre total des documents pertinents pour cette même requête. Par conséquent, le système de recherche d'information est efficace quand le Rappel et la Précision sont proches de 1. La F-mesure est la moyenne harmonique du Rappel et de la Précision : $F_{mesure} = 2\left[\frac{1}{R} + \frac{1}{P}\right]$

Il est clair que ces définitions de Rappel Précision F-mesure reposent sur des connaissances préalables sur la nature des documents (pertinente ou pas). Ce qui rend ces indices inapplicables dans le cas d'une classification non supervisée à cause de l'absence de classification de référence.

Ces indices ont cependant été adaptés au cas du clustering non supervisé [Lamirel et al, 2003]. Les mesures ne se font plus sur les documents mais sur les mots clés. L'idée est de mesurer l'homogénéité des classes en étudiant la distribution des mots clés à l'intérieur des classes. On introduit la notion de « mots propres » aux classes : chaque classe est caractérisée par un ensemble de mots clés dont les poids à l'intérieur de la classe par rapport à leurs poids dans la partition sont maximaux.

Plus explicitement, pour une partition $P = (C_1,...,C_k)$, on définit pour chaque classe C_i l'ensemble des mots propres suivant :

$$S_C = \left\{ p \in d, d \in C_i C \;\middle|\; \overline{W}_C^{\,p} = \max_{C' \in P}(\overline{W}_{C'}^{\,p}) \right\} \text{ avec } \overline{W}_C^{\,p} = \frac{\sum_{d \in C} W_d^p}{\sum_{C' \in P}\sum_{d \in C'} W_d^p}$$

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 la classe C à son poids total dans la partition. On définit aussi l'ensemble des classes propres comme suit : $\overline{P} = \{C \in P | S_C \neq \varnothing\}$

A partir de ces propriétés propres, les valeurs globales de Rappel et de Précision sont calculées de la manière suivante :

$$R = \frac{1}{|\overline{P}|} \sum_{C \in \overline{P}} \frac{1}{S_C} \sum_{p \in S_C} \frac{|c_p|}{|P_p|} \; ; \; P = \frac{1}{|\overline{P}|} \sum_{C \in \overline{P}} \frac{1}{S_C} \sum_{p \in S_C} \frac{|c_p|}{|c|}$$

c_p représente l'ensemble des documents de la classe C possédant la propriété *p* et P_p représente l'ensemble des documents de la partition *P* possédant la propriété *p*

Le Rappel mesure l'exhaustivité du contenu des classes. Il calcule, pour chaque mot propre associé à une classe donnée, la proportion des documents appartenant à la même classe et dans lesquels il apparaît par rapport au nombre total de documents contenant ce mot. Plus les classes présentent des mots propres exclusifs, plus la valeur globale du Rappel augmente corrélativement à la qualité du clustering.

La Précision permet de mesurer l'homogénéité du contenu des classes générées. Elle calcule pour un mot propre associé à une classe donnée, la proportion dans la classe des documents qui contiennent ce mot. Si les documents appartenant à une même classe ont tendance à avoir des descripteurs similaires, la précision globale augmente corrélativement à la qualité du clustering.

La F-mesure est définie de manière équivalente à celle de C.J.Van Rijsbergen.

Les valeurs globales de Rappel Précision et F-mesure varient entre 0 et 1. La qualité du clustering est meilleure quand R, P et F-mesure sont proches de 1. Cependant, le comportement de R et de P sont différents vis à vis du nombre de classes. En effet, plus le nombre de classes augmente, plus les effectifs à l'intérieur des classes diminuent, plus les valeurs $(|c_p| / |P_p|)$ diminuent et les valeurs $(|c_p| / |c|)$ augmentent. Par conséquent, R diminue avec le nombre de classe et P augmente avec le nombre de classes. Dans ce cas, un compromis possible entre R et P est de choisir la partition qui correspond à l'écart entre le Rappel et la Précision le plus faible. Le cas idéal est celui qui correspond à la coïncidence entre les deux valeurs.

5.2 Les résultats

Visualisons maintenant le comportement des indices Rappel Précision et F-mesure sur les résultats des méthodes de clustering (figure 5). Nous remarquons que les valeurs calculées ne présentent plus une similarité comme les autres indices de qualité traités ci-dessus. En effet, l'écart entre les méthodes se creuse aussi bien avec le Rappel qu'avec la Précision. Bien que ces indices aient des comportements monotones avec le nombre de classes (le Rappel diminue et la Précision augmente avec le nombre de classes), leur dépendance n'est pas aussi forte que celle de l'Inertie Intra classe et la Performance.

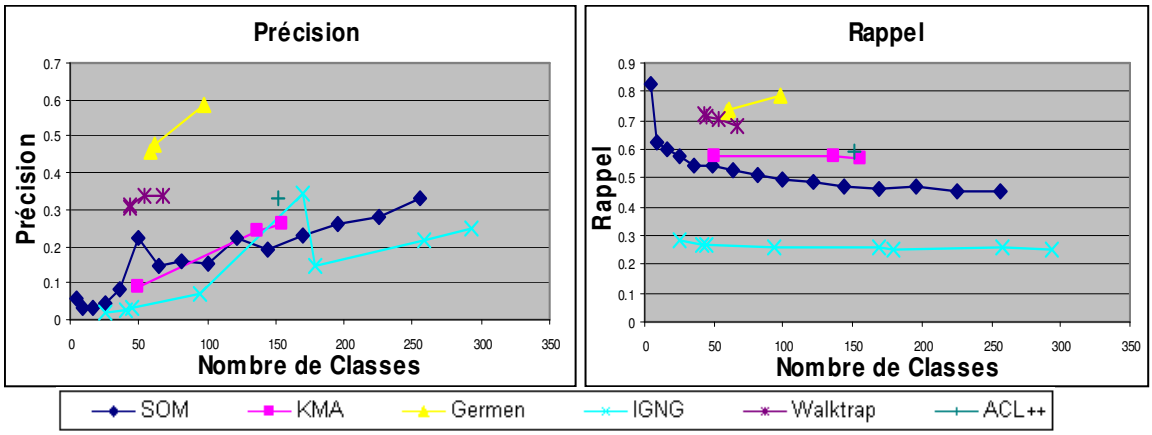

FIG. 5 – *Les valeurs des indices de Précision et Rappel en fonction du nombre de classes calculées pour les différentes méthodes de clustering*

On remarque aussi que Germen et Walktrap présentent des valeurs d'indices plus élevées. Les indices de Rappel Précision sont basés sur des moyennes calculées par classes. Ils se comportent comme des Macro-mesures et ne permettent donc pas d'identifier des classes hétérogènes comme celles produites par les méthodes Germen et Walktrap. Une alternative basée sur ces mesures est cependant possible. Elle consiste à calculer les valeurs de Rappel et de Précision globales en moyennant directement les valeurs de Rappel et de Précision des propriétés propres. Ces nouvelles mesures pourraient être considérées comme des Micro--mesures, indépendantes de la taille des classes.

6 Conclusions et Perspectives

L'évaluation d'un résultat de clustering reste une tâche très délicate. Des méthodes existent mais nos travaux montrent qu'elles ne sont pas satisfaisantes dans le cadre de l'analyse de données textuelles. Notre tentative d'élaborer de nouveaux indices basés sur la distribution des mots clés dans les classes n'a actuellement pas donné de bons résultats. Nous proposons d'axer nos travaux sur les trois points suivants :

- **Construction d'une classification de référence :** à partir des résultats de clustering obtenus sur notre corpus « La recherche en Lorraine », nous avons commencé la construction d'une classification « idéale » en corrigeant manuellement les erreurs constatées par des experts. Ce travail est long et fastidieux et il est vraisemblable qu'il n'y aura pas un résultat qui sera unique et indiscutable mais que l'on aura plutôt une certaine vision qui pourrait être différente en fonction des experts consultés. L'intérêt d'une telle approche est de garder un jeu de données réel, car des méthodes appliquées sur des jeux de données tests, plus ou moins simplistes, peuvent donner de bons résultats mais ne pas résister à l'épreuve des données réelles (données bruitées, incomplètes…).
- **Jeux de données artificielles :** une autre solution pourrait être de simuler artificiellement des données susceptibles d'être regroupées en un nombre déterminé

de classes. Le problème est alors de coller à la réalité de nos données particulières, à distributions des termes typiquement zipfiennes.

- **Construction d'indicateurs de qualité spécifiques :** cette alternative consiste à approfondir les approches basées sur les Micro-mesures telles que celles mentionnées à la fin de cet article. Selon nos expériences en cours, ces mesures semblent en effet présenter une alternative intéressante pour différencier les résultats de classification homogènes des résultats hétérogènes.

7 Références

Cadot M. (2006): Extraire et valider les relations complexes en sciences humaines : Statistique, motifs et règles d'association. Thèse Université de Franche-Comté, Besançon. 2006.

Cuxac P., Lelu A., Cadot M. (2009) : Suivi incrémental des évolutions dans une base d'information indexée : une boucle évaluation /correction pour le choix des algorithmes et des paramètres. 2ème conférence Internationale sur les systèmes d'informations et Intelligence Economique SIIE 2009, Hammamet Tunisie.

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

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

Jaccard P. (1901) Bulletin de la Société Vaudoise des Sciences Naturelles 37, 241-272.

Kohonen T. (1982) : Self-Organized Formation of Topologically Feature Maps. Department of technical physics, Helsinki University of technology, Espo, Finland.

Lamirel J.C., François C., Al Shehabi S., Hoffmann M. (2003): New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. In 9th International Conférence on Scientometrics an Informetrics - ISSI 2003, beiging, Chine.

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

Lelu A. (1994) : Clusters and Factors : neural algorithms for a noval representation of huge and highly multidimentionnal datasets. In New Approaches in Classification and Data Analysis, E.Diday, Y.Lechevallier al. editors, pp 241-248, Springer- Verlag, Berlin.

Lelu A., Cadot M. (2010) Slimming down a binary datatable: structural dimension and essential content. Soumis à COMPSTAT'2010, Paris, 22-27 Août 2010.

Lelu A.,Cuxac P.(2006): Classification dynamique d'un flux documentaire : une évaluation statique préalable de l'algorithme GERMEN. 8ème journée internationales d'Analyse statistique des Données textuelles, France.

Lewis D.D., Yang Y., Rose T., Li F., (2004) : RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research, 5:361-397, 2004.

Naïja Y., Sinaoui K.B. (2009) : A novel measure for validating clustering results applied to road traffic. In Proceedings of the Third international Workshop on Knowledge Discovery From Sensor Data (Paris, France, June 28 - 28, 2009). SensorKDD '09

Newman M.E.J., Girman M. (2004) : Finding an evaluating community structure in networks.Physical Review E, 69(6).

Pons P., Latapy M. (2006) : Computing communities in large networks using random walks. Journal of Graph Algorithms and Application.

Prudent Y., Ennaji A. (2004): Clustering incrémental pour un apprentissage distribué : vers un système évolutif et robuste. In Conférence CAp 2004,

Prudent Y., Ennaji A. (2005 a): A new learning algorithm for incremental self-organizing maps. TESANN'05 proceeding, European Symposium on Artificial Neural Networks, Brugs, Belgium,

Prudent Y., Ennaji A. (2005 b): An Incremental Growing Neural Gas learns Topologies. Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International

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.

Van Dongen S.M. (2000) : Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht

Van Rijsbergen C.J. (1979) : Information Retrieval. Butterworths, London.

Abstract: Our work on a new method for unsupervised classification (Germen) led us to question ourselves on the quality of results. The problem is to estimate whether a clustering method is 'better' than another for text data. Initially, after a state of the art of existing methods, we applied some quality indices to clustering results from our Germen algorithm and other algorithms commonly used. These quality indices did not allow us to select the best partition, so we have developed a new series of indices based on the distribution of keywords. We present and discuss the results obtained and reflections initiated to evolve the evaluation of unsupervised text classification.