Regroupements non-disjoints de mots pour la classification de documents

Guillaume Cleuziou

LIFO, Laboratoire d'Informatique Fondamentale d'Orléans
Université d'Orléans
BP 6759 - 45067 ORLEANS Cedex 2
guillaume.cleuziou@lifo.univ-orleans.fr

RÉSUMÉ. *La classification automatique de documents est un domaine d'étude en plein essor dans le domaine du Traitement et de la Recherche d'Information (RI). Dans un cadre supervisé, il s'agit alors d'entraîner un modèle de classifieur sur un corpus de documents étiquetés. La difficulté majeure consiste à représenter les documents par un nombre limité et suffisant d'attributs. Dans cet article, nous proposons une méthode de regroupement de mots, basée sur l'algorithme PoBOC (Pole-Based Overlapping Clustering) autorisant les recouvrements entre les groupes. Ainsi, chaque mot initial peut appartenir à un ou plusieurs attributs terminaux. Les expérimentations menées sur le corpus Reuters-21578 ont permis de montrer que cette méthode de regroupements non-disjoints induit, sous de bonnes conditions, une amélioration de la précision du classifieur.*

ABSTRACT. *Nowadays, automatic document categorization is an important challenge in the Information Retrieval (IR) and Processing field. From a supervised point of view, this task consists in training a categorization model (classifier) on a corpus of documents. The major problem concerns the representation of the documents in a feature space of reasonable dimension. In this paper we propose a new method to cluster words in overlapping groups. This approach is based on the PoBOC (Pole-Based Overlapping Clustering) algorithm which allows a word to appear in one or several features. Experiments on the Reuters-21578 corpus show that overlapping features lead to an improvement in classification accuracy, on well defined conditions.*

MOTS-CLÉS : *Classification de documents, regroupements non-disjoints, regroupement distributionnel, regroupement d'attributs, apprentissage supervisé.*

KEYWORDS : *Document classification, overlapping clustering, distributional clustering, feature clustering, supervised learning.*

1. Introduction

La classification de documents est un domaine d'étude en plein essor en raison de la quantité d'information qui transite, notamment sur Internet, et de la valeur stratégique qu'elle revêt. La classification peut consister en un processus supervisé ou non-supervisé. Dans le cas de la classification non-supervisé, il s'agit de proposer une organisation des objets (ici des documents) en classes, selon un critère de similarité ou dissimilarité à partir d'une description sur ces objets (Slonim *et al.*, 2000). L'une des difficultés majeures de l'apprentissage non-supervisé concerne l'évaluation des classes constituées. En revanche dans un cadre supervisé, il s'agit d'apprendre un modèle, ou classifieur, à partir d'un ensemble d'entraînement composé de couples *(objet, classe)* puis de tester la qualité du classifieur appris sur un ensemble d'objets tests. C'est dans ce contexte que se situe notre étude.

Afin d'apprendre à classer des objets par rapport à un ensemble cible de classes, ces derniers doivent être décrits selon des attributs (ou traits) « pertinents » ; or la construction d'un ensemble de tels attributs est une tâche difficile, notamment dans l'application à la classification de documents. En effet, lorsqu'il s'agit de classer automatiquement les documents par thématiques, l'analyse sémantique semble être, sinon l'unique, du moins la principale caractérisation possible des documents. Ainsi chaque document est perçu comme un "sac de mots" et l'ensemble du vocabulaire contenu dans les documents comme l'ensemble des attributs possibles. Cependant, la quantité de vocabulaire, son caractère redondant dans l'influence des mots pour la classification, et l'éparsité de la matrice résultante *(documents × mots)* sont autant de critères en faveur d'une réduction de la dimension de l'espace de description (Aas *et al.*, 1999).

Plusieurs approches ont été proposées dans ce sens : la sélection des attributs pertinents en mesurant l' « intérêt » de chaque mot afin de supprimer ceux qui apportent peu d'information (gain d'information (Yang *et al.*, 1997), test du χ^2 (Liu *et al.*, 1995), etc.) ; le re-paramétrage des attributs pour en définir de nouveaux à partir de combinaisons et transformations des traits initiaux (LSI (Deerwester *et al.*, 1990)) ; enfin le regroupement des attributs, permettant de considérer les mots ayant un rôle similaire dans la classification comme un seul attribut (Baker *et al.*, 1998). Cette dernière méthode, comparativement aux deux premières, permet de compresser l'espace des attributs de manière plus agressive tout en conservant un taux de classification[1] important. Ces travaux sont basés sur l'analyse distributionnelle des mots proposée dans (Pereira *et al.*, 1993) puis reprise par (Baker *et al.*, 1998) dans le cadre de la classification de documents. Depuis, plusieurs études ont été menées proposant des algorithmes de regroupement plus adaptés aux données traitées afin d'améliorer la qualité de l'ensemble des nouveaux

[1] Le taux de classification est donné par le rapport entre le nombre de documents dont la classe attribuée est correcte et le nombre total de documents à classer.

attributs constitués (Slonim *et al.*, 2000), (Dhillon *et al.*, 2002). Un point commun à ces algorithmes est de proposer une organisation des mots en groupes disjoints, c'est-à-dire que chaque mot ne peut appartenir qu'à un seul groupe final. Si ce choix de méthode peut se justifier pour des raisons pratiques[2], cette stratégie n'est pas adaptée dans le contexte du regroupement de mots.

L'étude que nous proposons ici consiste à introduire le concept de « groupes non-disjoints » de mots pour définir le nouvel ensemble d'attributs dans la perspective de classification de documents. Nous étudierons alors les conditions dans lesquelles ces recouvrements peuvent améliorer la qualité du classifieur, puis nous en observerons l'impact qualitatif sur la classification. Pour cela nous baserons nos expérimentations sur le corpus Reuters-21578.

L'article est organisé comme suit : nous proposons en section 2, une présentation générale du classifieur naïf de Bayes et définissons une mesure de dissimilarité entre mots ; la section suivante est dédiée à la présentation de l'algorithme de regroupements non-disjoints (PoBOC) ; le chapitre suivant permet dans un premier temps de cibler les hypothèses sous lesquelles les recouvrements induisent un gain d'information, puis dans un second temps de mesurer l'amélioration apportée par un tel classifieur comparativement à l'approche de (Baker *et al.*, 1998). Enfin, la section 5 dresse un bilan de cette expérience avant de proposer de nouvelles perspectives à ce domaine d'étude.

2. Classifieur naïf de Bayes et analyse distributionnelle

2.1. Approche bayesienne pour la classification des documents

Le classifieur naïf de Bayes, traditionnellement utilisé pour la classification de documents en raison de ses performances reconnues dans ce domaine (Lewis, 1991), suppose l'existence d'un modèle de génération d'un document à partir duquel on peut déduire la ou les classes les plus probables d'appartenance du document. Les paramètres du modèle sont estimés sur un corpus d'entraînement.

Soit $D=\{d_1,...,d_n\}$ l'ensemble des documents constituant le corpus d'entraînement, chacun étant étiqueté par une ou plusieurs classes de $C=\{c_1,...,c_m\}$. L'ensemble du vocabulaire présent dans D est noté $W=\{w_1,...,w_p\}$. On cherche à apprendre un ensemble $\{\theta\}$ de paramètres sur D tel que la probabilité *a priori* $P(c_j/d;\theta)$ soit élevée pour la ou les classes associées au document d. Cette probabilité *a priori* est définie par [1].

$$P(c_j/d;\theta) = \frac{P(c_j/\theta).P(d/c_j;\theta)}{P(d/\theta)} \quad [1]$$

[2] Il existe peu d'algorithmes autorisant les recouvrements entre les groupes (cf. section 3.2.).

$P(c_j/\theta)$ correspond alors au rapport du nombre de documents étiquetés c_j dans D sur le nombre total de documents dans D ; $P(d/c_j;\theta)$ est calculée, sous l'hypothèse naïve d'indépendance entre les mots d'un même document, par [2].

$$P(d/c_j;\theta)=P(|d|)\prod_{w_t\in d}P(w_t/c_j;\theta) \quad [2]$$

Enfin, la probabilité d'un document $P(d/\theta)$ est donnée par [3].

$$P(d/\theta)=\sum_{k=1}^{m}P(c_k/\theta).P(d/c_k;\theta) \quad [3]$$

La probabilité d'apparition d'un mot sachant la classe $P(w/c_j;\theta)$ est utilisée dans [2] et [3]. Cette quantité est estimée par la règle de succession de Laplace ([4]).

$$P(w_t/c_j;\theta)=\frac{1+\sum_{d\in c_j}n(w_t,d)}{|W|+\sum_{w\in W}\sum_{d\in c_j}n(w,d)} \quad [4]$$

Dans [4], *n(w,d)* correspond au nombre d'occurrences du mot *w* dans le document *d*. Par simplifications, le classifieur naïf de Bayes se résume à l'équation [5], dans le cas classique où les attributs sont les mots.

$$c^*(d)=\underset{c_j\in C}{\operatorname{argmax}}\left(\frac{\log(P(c_j/\theta))}{|d|}+\sum_{w_t\in W}P(w_t/d).\log(P(w_t/c_j;\theta))\right) \quad [5]$$

Dans cet article, nous redéfinissons les attributs comme des groupes de mots, notés $\hat{W}=\{\hat{w}_1,\ldots,\hat{w}_l\}$. Les formules [4] et [5] deviennent respectivement [6] et [7].

$$P(\hat{w}_s/c_j;\theta)=\frac{1+\sum_{d\in c_j}n(\hat{w}_s,d)}{l+\sum_{\hat{w}\in\hat{W}}\sum_{d\in c_j}n(\hat{w},d)} \quad [6]$$

$$c^*(d)=\underset{c_j\in C}{\operatorname{argmax}}\left(\frac{\log(P(c_j/\theta))}{|d|}+\sum_{\hat{w}_s\in\hat{W}}P(\hat{w}_s/d).\log(P(\hat{w}_s/c_j;\theta))\right) \quad [7]$$

Dans cette dernière formule ([7]), la probabilité d'un groupe de mots sachant le document $P(\hat{w}_s/d)$ est définie par [8].

$$P(\hat{w}_s/d)=\frac{\sum_{w\in\hat{w}_s}n(w,d)}{|d|} \quad [8]$$

Pour plus de détails sur l'ensemble des formalismes et des transformations présentées dans cette section, nous invitons le lecteur à se reporter à l'étude proposée par (Dhillon *et al.*, 2003).

2.2. Définition de la similarité entre mots

De nombreuses mesures de similarité ou dissimilarité entre les mots ont été définies jusqu'à présent. Citons par exemple quelques mesures fondées sur les cooccurrences de mots dans les corpus : la mesure d'*Information Mutuelle* (Fano, 1961), le *Rapport d'Association* (Church *et al.*, 1989) ainsi que d'autres coefficients tels que le *coefficient de Dice* (Sneath *et al.*, 1973) ou la *mesure de Jaccard* (Grefenstette, 1994). D'autres approches permettent d'intégrer des connaissances extérieures au corpus, qu'il s'agisse de thésaurus ou de listes thématiques (Resnik, 1995) ou encore d'utiliser le Web comme ressource linguistique (Turney, 2001), (Clavier *et al.*, 2002). Enfin, dans le cadre supervisé de la classification de documents, (Baker *et al.*, 1998) s'appuient sur l'approche distributionnelle afin de définir une mesure de dissimilarité relativement aux classes cibles que l'on cherche à apprendre. Chaque mot est alors défini par sa distribution sur la variable de classe $P(C/w)$ et la dissimilarité entre deux distributions est donnée par « *la divergence de Kullback-Leibler (KL) à la moyenne* ». Formellement, cette mesure est définie par [9].

$$\begin{aligned} d(w_t,w_s) = P(w_t).D(P(C/w_t) \,\|\, P(C/w_t \vee w_s)) \\ + P(w_s).D(P(C/w_s) \,\|\, P(C/w_t \vee w_s)) \end{aligned} \quad [9]$$

où $D(P(C/w_t) \,\|\, P(C/w_s))$ représente la divergence de KL[3] entre les distributions des deux mots w_t et w_s. $P(C/w_t \vee w_s)$ est donnée par [10].

$$\frac{P(w_t)}{P(w_t)+P(w_s)} P(C/w_t) + \frac{P(w_s)}{P(w_t)+P(w_s)} P(C/w_s) \quad [10]$$

Cette mesure de dissimilarité permet ainsi de comparer deux mots relativement à leur distribution sur la variable de classe.

---

[3] $D(P(C/w_t) \,\|\, P(C/w_s)) = \sum_{k=1}^{m} P(c_k/w_t) \log\left(\frac{P(c_k/w_t)}{P(c_k/w_s)}\right)$

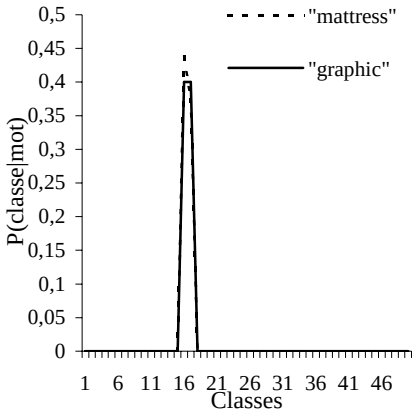


Figure 1. *Distributions de deux mots fortement similaires sur 50 classes de Reuters-21578*

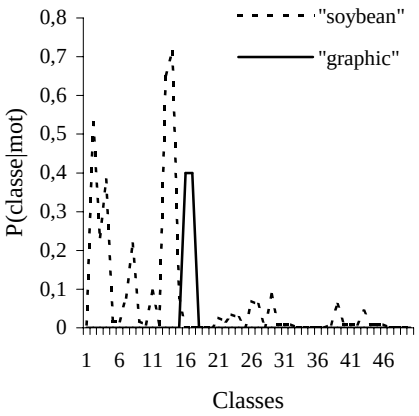


Figure 2. *Distributions de deux mots fortement dissimilaires sur 50 classes de Reuters-21578*

Nous choisissons, par exemple, un couple de mots extraits du corpus Reuters-21578, dont la dissimilarité est parmi les plus faibles. Les mots *« graphic »* et *« mattress »* n'entretiennent pas, à première vue, de relation sémantique, cependant leur distribution respective par rapport aux classes cibles (cf. figure 1) indique que ces deux mots jouent un rôle commun, ce qui explique leur proximité statistique. De la même manière, on observe que les deux mots *« graphic »* et *« soybean »* disposent d'une forte dissimilarité, que l'on peut à nouveau constater sur leur distribution (cf. figure 2).

3. Regroupement et recouvrements d'attributs

3.1. Les recouvrements entre attributs : motivations

Le regroupement des attributs (*feature clustering*) a pour objectif de regrouper sous un seul attribut, un ensemble de plusieurs traits initiaux ayant un même rôle dans la tâche de classification. Autrement dit, dans l'application aux documents, il s'agit en quelque sorte de définir de nouveaux mots « fantômes » représentant plusieurs mots attestés, extraits des documents et distribués de manières assez similaires sur les documents et donc, indirectement, sur les classes de documents.

Dans le cadre du regroupement sémantique, les algorithmes de regroupement traditionnels (*e.g.* l'algorithme des *k*-moyennes) sont généralement peu appréciés car ils ne permettent pas de représenter l'ensemble des relations existant entre les mots et entre les concepts. D'autre part, les algorithmes de regroupement flou (*e.g.* la version floue de l'algorithme des *k*-moyennes) tiennent compte des caractéristiques précédentes sans toutefois proposer de représentations exploitables.

Cette remarque, portant sur le regroupement sémantique, est également valable pour le regroupement distributionnel.

Considérons, un ensemble d'entraînement composé de 4 documents $\{d_1,d_2,d_3,d_4\}$ desquels sont extraits les 5 attributs (mots) suivants $V= \{w_1,w_2,w_3,w_4,w_5\}$. Les documents sont caractérisés par la matrice *(document × mot)* suivante,

$$M=\begin{pmatrix}11000\\11100\\00111\\00011\end{pmatrix}$$

telle que $M(i,j)$ correspond au nombre d'apparitions de w_j dans d_i. Supposons que d_1 et d_2 soient étiquetés c_1 alors que d_3 et d_4 appartiennent à la classe c_2. Les deux mots w_1 et w_2 (respectivement w_4 et w_5) sont distribués identiquement sur les deux classes c_1 et c_2, ils ont donc une dissimilarité égale à *0* par la mesure définie en [9]. Si l'on cherche à former deux groupes sur V, on obtiendra alors les deux sous-ensembles $W_1=\{w_1, w_2\}$ et $W_2=\{w_4, w_5\}$; le problème se pose alors pour w_3 de similarité identique avec les deux groupes précédents. La formation de groupes disjoints conduit à affecter arbitrairement w_3 à l'un des deux groupes, par exemple W_1. On obtient alors les probabilités à priori suivantes : $P(c_1|W_1)= 1 ; P(c_1|W_2)= 0 ; P(c_2|W_1)= 0.5 ; P(c_2|W_2)= 1.$ Ce modèle favorise le classement d'un nouveau document test $d'=(0,0,1,0,0)$ dans c_1, de même que l'affectation de w_3 au groupe W_2 aurait influencé le classement en faveur de c_2. En revanche, le fait d'autoriser w_3 à appartenir aux deux groupes W_1 et W_2 permet de conserver l'ambiguïté de classement du document test, conformément aux données d'apprentissage.

Plus formellement, (Diday, 1984) montre, dans le cas hiérarchique, qu'une représentation en groupes non-disjoints est plus fidèle aux données initiales qu'une organisation en classes disjointes.

3.2. Algorithme de regroupement avec recouvrements (PoBOC)

Tout d'abord, il convient de préciser quelques aspects terminologiques dans le domaine du regroupement (ou *clustering*). Le regroupement « flou » (ou *fuzzy clustering*) fait référence aux algorithmes proposant en sortie, un ensemble de foyers ainsi qu'une matrice d'appartenance floue ; chaque objet appartient donc plus ou moins à chaque classe. Ce type de représentation est très riche mais peu exploitable sans post-traitement. L'approche de regroupement flou la plus célèbre est l'algorithme des *k*-moyennes flou (ou *fuzzy-c-means)* (MacQueen, 1967). On oppose à cette génération d'algorithmes, les méthodes dites de regroupement « dur » (ou *hard clustering*), pour lesquelles chaque objet est affecté à un unique cluster final. Les principaux algorithmes dans ce domaine sont les algorithmes hiérarchiques (SAHN (Sneath *et al.*, 1973)) ou de partitionnement (*k*-moyennes (MacQueen, 1967)). Pour plus de précisions sur ces techniques de clustering, nous renvoyons le lecteur aux travaux de synthèse proposés par (Jain *et al*., 1999).

L'approche que nous envisageons dans cette étude consiste à construire des clusters non-disjoints, où chaque objet peut appartenir à un ou plusieurs clusters. Cette technique se situe entre les deux visions précédentes et est appelée : regroupement avec recouvrements ou regroupement non-disjoint. Nous éviterons d'utiliser les termes « regroupement flou » et « *soft clustering* », pouvant prêter à confusion.

Comparativement aux regroupements durs et flous, il existe relativement peu d'algorithmes de regroupement avec recouvrements. Une méthode générale est celle des « classes empiétantes » ou « pyramides » proposée par (Diday, 1984). Enfin, d'autres techniques de ce genre ont été envisagées récemment dans le cadre d'applications très précises comme par exemple l'algorithme WBSC (Lin *et al.*, 2001) pour regrouper des documents de façon non-supervisée ou plus récemment encore la catégorisation floue de pages HTML par l'algorithme des *k-means axiales* (Lelu *et al.*, 1999). Nous choisissons de nous inspirer ici de l'algorithme PoBOC (*Pole-Based Overlapping Clustering*) présenté dans (Cleuziou *et al*., 2003).

Soient :	$V=\{w_1,\ldots,w_n\}$ l'ensemble des objets à regrouper et M la matrice de similarités sur $V\times V$
Initialisation :	Construire un graphe de similarités $G_M(V)$,
Etape 1 :	Construire l'ensemble $\mathcal{P}$ des pôles $\{P_1,\ldots,P_l\}$ par la recherche de cliques dans G_M,
Etape 2 :	Construire la matrice U des appartenances sur $\mathcal{P}\times V$ telle que :
	$U(P_i,w_j)$ est la similarité moyenne de w_j avec chaque composant du pôle P_i,
Etape 3 :	Pour chaque objet $w_i \in V$, affecter w_i à un ou plusieurs pôles de $\mathcal{P}$,
Etape 4 :	Construire l'arbre hiérarchique partir des clusters obtenus.

Algorithme 1. *Présentation générale de l'algorithme PoBOC*

Cet algorithme (cf. Algorithme 1) se base sur la construction de « pôles » dans le graphe des similarités entre les objets, suivie d'une étape de « multi-affectations » des objets aux pôles. Finalement, les clusters obtenus sont représentés sous-forme d'une hiérarchie de concepts. PoBOC présente alors les avantages suivants : (1) le nombre de clusters à constituer n'est pas donné à priori, mais est déterminé automatiquement en relation avec la configuration des objets ; (2) le résultat final ne dépend d'aucune initialisation (contrairement aux approches du type *k*-moyennes) ; (3) l'algorithme est général et s'applique à des objets de toutes natures sous-réserve de la définition d'une mesure de similarité ; (4) les recouvrements entre clusters ne sont pas soumis aux contraintes d'ordonnancement des clusters, contrairement à l'approche pyramidale pour laquelle chaque cluster possède une intersection avec au plus deux autres clusters.

La principale difficulté dans l'utilisation du regroupement de données textuelles est la quantité très importante de données à traiter. Pour contourner ce problème on peut avoir recours à des techniques telles que l' « échantillonnage », qui consiste à traiter un sous-ensemble représentatif des données plutôt que l'ensemble complet. Nous verrons dans la suite que PoBOC se prête bien à ce genre d'adaptation.

4. Traitement du corpus et résultats expérimentaux

4.1. Présentation et traitement du corpus Reuters-21578

Le corpus Reuters-21578[4] est composé de 21578 documents extraits du journal « Reuters » en 1987. Ce corpus est souvent utilisé comme base de comparaison entre les différents outils de classification de documents. D'autre part, on retiendra que le Reuters-21578 est souvent qualifié de « corpus difficile » pour des traitements complexes.

Nous utilisons la collection modifiée « ModApte » (Apte *et al.*, 1994) pour en extraire deux sous-corpus : l'ensemble d'apprentissage est alors constitué de 9603 articles et celui de test de 3299 articles. Chacun de ces articles est étiqueté par une ou plusieurs classes parmi un ensemble de 114 classes au total.

Les expérimentations proposées dans la suite de cet article, consistent à apprendre sur le corpus de 9603 articles, puis à tester le classifieur sur les 3299 articles tests. 19646 lemmes sont extraits du corpus d'entraînement par suppression des mots vides[5] et des mots de taille inférieure à 3, puis par lemmatisation en utilisant le « Porter Stemmer » (Porter, 1980).

Cet ensemble de lemmes servira alors de base de description pour les documents d'entraînement et les documents tests à classer. La performance globale d'un classifieur est évaluée dans la suite de l'étude, par le rapport du nombre total de documents tests correctement classés sur le nombre total de documents tests. Un document test est correctement classé, si la classe prédite par le classifieur[6] appartient à l'ensemble des classes proposées pour ce document.

4.2. Conditions d'amélioration de la performance du classifieur par les recouvrements entre attributs

L'idée de diminuer l'espace de description des documents en regroupant les attributs (mots) en groupes non-disjoints semble raisonnable compte tenu des arguments avancés en section 3.1. Cependant, cette intuition reste à démontrer empiriquement et nous amène à réfléchir sur les conditions dans lesquelles ces

[4] Ce corpus est disponible à l'adresse : http://research.att.com/~lewis/reuters21578.html.
[5] Utilisation d'une liste de mots vides.
[6] Classe pour laquelle le classifieur obtient le meilleur score.

recouvrements entre attributs seraient bénéfiques pour la tâche plus générale de classification. Les recouvrements entre attributs engendrent-ils effectivement un gain ? Si oui, est-ce indépendant du nombre d'attributs générés et/ou de la quantité de recouvrements autorisés ?

Soient :	$V=\{w_1,...,w_n\}$ l'ensemble des objets à regrouper, M la matrice de similarité sur $V \times V$ déduite de [6], et k le nombre de groupes,
Etape 1 :	Ordonner V par information mutuelle[7] décroissante avec la variable de classe C, on obtient $V=\{w_1',...,w_n'\}$
Etape 2 :	Soit $\mathcal{P} = \{\{w_1'\},\{w_2'\},...,\{w_n'\}\}$ l'ensemble des k pôles, Pour chaque $w_i \in \{w_{k+1}',...,w_n'\}$ - calculer l'appartenance de w_i à chaque pôle de $\mathcal{P}$ - affecter w_i à un ou plusieurs pôles

Algorithme 2. *Algorithme de regroupement simplifié*

Afin de répondre à ces questions, nous entreprenons dans un premier temps d'utiliser l'algorithme PoBOC réduit aux étapes 2 et 3 sur un échantillon représentatif de l'ensemble des 19646 lemmes extraits (cf. Algorithme 2).

Nous rappelons également la méthode d'affectation des objets aux pôles par les définitions suivantes :

Définition 3.2.1.

Soient V un ensemble d'objets à traiter, $\mathcal{P}=\{P_1,...,P_l\}$ un ensemble de l pôles sur V et M une matrice de similarité sur l'ensemble V. La valeur d'appartenance d'un objet $w \in V$ à un pôle $P_i \in \mathcal{P}$ est définie par [11] :

$$u(P_i,w)=\frac{1}{|P_i|}\sum_{w_i \in P_i} M(w,w_i) \qquad [11]$$

Définition 3.2.2.

Soient V un ensemble d'objets à traiter, $\mathcal{P}=\{P_1,...,P_l\}$ un ensemble de l pôles sur V et M une matrice de similarité sur l'ensemble V. Pour tout objet $w_k \in V$, les pôles

---

[7] $IM(w,C)=P(w).\sum_{c_j \in C} P(c_j/w).\log\frac{P(c_j/w)}{P(c_j)}$

de $\mathcal{P}$ sont ordonnés relativement à w_k. P_{ki} correspond alors au $j^{ième}$ plus « proche » pôle de w_k tel que $i<j \Rightarrow u(P_{kj},w_k)<u(P\ P_{ki},w_k)$. L'objet w_k est affecté au pôle $P_{k,j}$ si et seulement si, l'une des trois propriétés suivantes est vérifiée :

i) $j=1$ *(P_{kj} est le pôle le plus proche de w_k)*

ii) $1<j<l,\ u(P_{k,j},w_k) \geq m.(\ u(P_{k,j-1},w_k) - u(P_{k,j+1},w_k)) + u(P_{k,j+1},w_k)$ *et ii) est vérifiée* $\forall j'<j$

iii) $j=l,\ u(P_{k,j},w_k) \geq m.(\ u(P_{k,j-1},w_k))$ *et ii) est vérifiée* $\forall j'<j$

Dans cette dernière définition, le paramètre *m* sera considérer comme *fuzzifieur*. Lorsque *m* vaut *1*, un objet est affecté uniquement au pôle qui lui est le plus proche. Quand *m* vaut *0*, l'objet est affecté à chacun des pôles.

Une première expérimentation consiste à comparer la qualité des attributs avec ou sans recouvrements ($m<1$ ou $m=1$) en observant la performance du classifieur sur le corpus test. Le graphique (figure 3) présente le gain (ou perte) de performance du classifieur induit par l'utilisation d'attributs avec recouvrements comparativement au classifieur « témoin », sans recouvrements. Le choix de $m=0.7$ est guidé par le taux de recouvrements[8] aux alentours de *25%*, jugé raisonnable. On conclut alors à la confirmation de l'hypothèse principale à savoir que les recouvrements entre les groupes de mots peuvent, sous certaines conditions, améliorer la qualité des attributs terminaux pour la classification de documents.

Une première condition à vérifier concerne le nombre d'attributs choisis. Nous remarquons, toujours figure 3, que le gain par rapport à un regroupement disjoint augmente entre 5 et 30 clusters puis chute brusquement à partir de 40 attributs. Finalement, le recouvrement entre les groupes de mots n'est pas bénéfique lorsque le nombre d'attributs est trop important (ici à partir de 50 attributs). Ce phénomène pourrait s'expliquer comme suit : il existerait une limite au-delà de laquelle les attributs sont suffisamment précis et homogènes, car de tailles plus petites, pour ne pas nécessiter de recouvrements, justement utilisés pour affiner les attributs. L'intérêt du regroupement d'attributs étant de réduire le plus possible la dimension de l'espace de description des documents, tout en conservant un classifieur de bonne qualité, il est justifié de s'attacher à améliorer la performance du classifieur dans un espace à faible dimension.

[8] Le taux de recouvrement entre les groupes est déterminé par le quotient du nombre d'affectations par le nombre total d'objets moins 1.

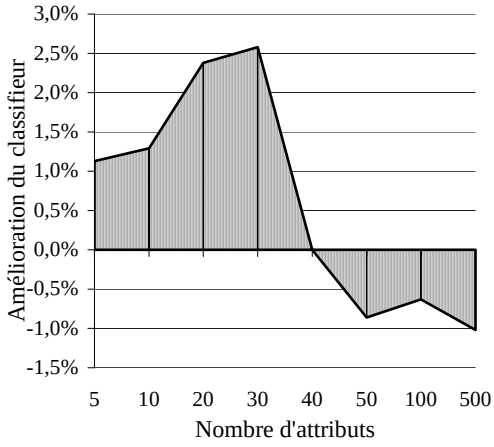
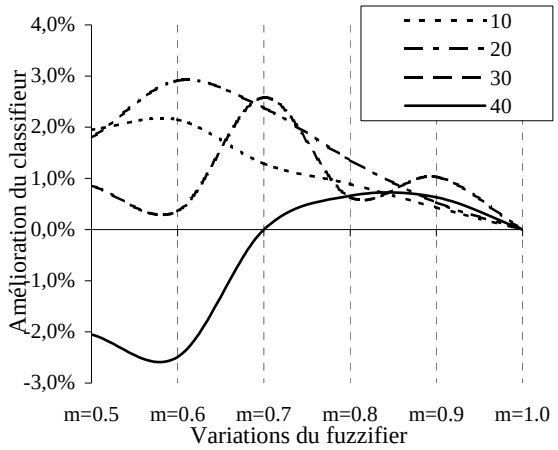


Figure 3. *Gain induit par les recouvrements entre attributs (m=0.7), par rapport au regroupemnt disjoint d'attributs (m=1).*

Une première condition à vérifier concerne le nombre d'attributs choisis. Nous remarquons, toujours figure 3, que le gain par rapport à un regroupement disjoint augmente entre 5 et 30 clusters puis chute brusquement à partir de 40 attributs. Finalement, le recouvrement entre les groupes de mots n'est pas bénéfique lorsque le nombre d'attributs est trop important (ici à partir de 50 attributs). Ce phénomène pourrait s'expliquer comme suit : il existerait une limite au-delà de laquelle les attributs sont suffisamment précis et homogènes, car de tailles plus petites, pour ne pas nécessiter de recouvrements, justement utilisés pour affiner les attributs. L'intérêt du regroupement d'attributs étant de réduire le plus possible la dimension de l'espace de description des documents, tout en conservant un classifieur de bonne qualité, il est justifié de s'attacher à améliorer la performance du classifieur dans un espace à faible dimension.

Dans un second temps, nous cherchons à évaluer l'impact du paramètre *m* sur l'amélioration observée. La figure 4 rend compte de la variation de qualité des attributs suivant le nombre d'attributs (10 à 40) et l'importance du recouvrement autorisé. Le tableau situé sous le graphique de la figurer 4, indique le taux moyen de recouvrement entre groupes pour chaque valeur du paramètre *m*. Lorsque *m* varie entre *0.5* et *1.0*, le taux de recouvrement diminue de *92%* à *0%.* On observe la déformation de la courbe de gain lorsque le nombre d'attributs atteint la valeur 30 pour passer en négatif à partir de 40 clusters. Enfin, cette étude permet de confirmer le choix du paramètre *m=0.7*, comme compromis entre le manque et l'excès de recouvrements.



m	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*	*1.0*
% rec. moyen	*92,26%*	*56,43%*	*26,46%*	*19,57%*	*13,67%*	*0,00%*

Figure 4. *Gain induit par les recouvrements entre attributs, par rapport au regroupement disjoint. Variation du taux de recouvrement autorisé.*

Nous avons donc confirmé l'idée que l'utilisation d'algorithmes de regroupements non-disjoints permet d'améliorer la qualité des attributs ainsi générés. Ce gain est cependant vérifié sous deux conditions : un nombre d'attributs et un taux de recouvrements limités.

4.3. Comparaisons entre l'algorithme PoBOC et le regroupement Distributionnel Agglomératif (ADC)

Les études menées précédemment permettent d'évaluer l'impact des recouvrements entre groupes, sur une même méthode de regroupement. Nous souhaitons à présent comparer la qualité respective des groupes obtenus par PoBOC et celle des attributs obtenus par les approches existantes. Le tableau 1 présente une étude comparative donnant le taux de bonne classification induit par trois méthodes de réduction du nombre d'attributs :

- Sélection des attributs par Information Mutuelle (IM) : cette technique consiste à ordonner de façon décroissante, chaque attribut (ici les mots ou lemmes) par Information Mutuelle avec la variable de classe (Yang *et al.*, 1997). Pour une réduction à *k* attributs, seuls les *k* premiers mots sont utilisés pour classer les documents tests.

- Regroupement des attributs par ADC : cette technique est celle proposée par (Baker *et al.*,1998). L'algorithme procède par fusions successives des deux plus proches groupes, en commençant par les *k* clusters réduits chacun à un mot (les *k*

mots d'Info. Mutuelle maximale) puis en ré-injectant un nouveau singleton après chaque fusion.

- Regroupement des attributs par PoBOC : nous utilisons ici l'algorithme PoBOC complet. Pour N mots de départ, PoBOC construit un nombre $k<<N$ de pôles et donc de groupes terminaux. Pour les besoins de l'étude, nous avons observé la variation du nombre de pôles formés relativement au nombre de mots fournis afin de choisir, pour chaque valeur k désirée, le nombre minimum de mots nécessaires afin d'obtenir k clusters en sortie. Les pôles sont cette fois définis par un ensemble de mots plutôt que par un seul comme c'était le cas précédemment. Ceci permet d'améliorer considérablement la qualité des pôles et donc du regroupement final.

Méthode	*5 attributs*	*10 attributs*	*20 attributs*	*30 attributs*	*40 attributs*	*50 attributs*
IM	33,6%	37,6%	46,3%	50,0%	62,0%	63,8%
ADC	57,1%	70,6%	74,5%	77,6%	78,1%	80,3%
PoBOC	62,7%	66,8%	75,4%	78,2%	79,6%	79,8%

Tableau1. *Performances du classifieur suivant le nombre d'attributs. Comparaisons entre trois méthodes de réduction.*

L'expérience menée confirme l'hypothèse formulée sur l'intérêt de regrouper les attributs plutôt que de les sélectionner. Les résultats obtenus par les deux approches de regroupement sont nettement meilleurs que pour la méthode de sélection des attributs par Information Mutuelle. Enfin et surtout, on observe une amélioration de la performance du classifieur basé sur les attributs non-disjoints par rapport au classifieur basé sur les attributs disjoints, notamment dans l'intervalle de 20 à 50 attributs. Cette dernière observation vient alors conforter la thèse initiale selon laquelle « le regroupement d'attributs par un algorithme autorisant les recouvrements permet d'améliorer la qualité des attributs dans une perspective de classification supervisée de documents ».

5. Conclusion et perspectives

Ce travail a permis de proposer une amélioration des techniques actuelles de regroupement d'attributs dans le carde de la classification supervisée de documents. En nous appuyant sur les travaux existant dans le domaine de l'analyse distributionnelle, nous postulons que le regroupement d'attributs autorisant les intersections (ou recouvrements) permet d'améliorer la qualité et la précision de ces nouveaux critères de description pour les documents. Les expérimentations, effectuées sur le corpus Reuters-21578, ont montré dans un premier temps le gain induit par les recouvrements sous de bonnes conditions relatives à l'importance de la réduction de l'espace de description et la maîtrise de la proportion d'intersections.

Enfin, nous avons observé que les attributs construits à l'aide de l'algorithme PoBOC induisent un classifieur de précision meilleure que d'autres approches telles que le regroupement d'attributs par agglomération ou encore la méthode de sélection des attributs par le critère d'Information Mutuelle.

Ces résultats nous encouragent à poursuivre cette étude en expérimentant cette technique sur d'autres corpus de tailles et de spécificités variées, afin de confirmer les conclusions sur le gain induit par les recouvrements entre attributs, et de mesurer la dépendance des paramètres (fuzzifieur, nombre limite d'attributs,...) au corpus utilisé. Nous tâcherons également par la suite de tenir compte de la particularité du corpus Reuters-21578 dans lequel chaque document est étiqueté par plusieurs classes. De plus, nous envisageons de comparer l'algorithme PoBOC avec d'autres algorithmes tels que les pyramides ou une adaptation des *k*-moyennes aux classes recouvrantes. Enfin, ce travail pourra être complété par une comparaison plus avancée avec d'autres techniques de réduction telles que l'approche LSI et de classification (Support Vector Machine, Rocchio...).

6. Bibliographie

Aas L., Eikvil L., Text categorization: A survey, Rapport n°941, 1999, Norwegian Compution Center.

Apte C. Damerau F., Weiss S., « Automated Learning of Decision Rules for Text Categorization », *Information Systems*, vol. 12, n°3, 1994, p.233-251.

Baker L., McCallum A., « Distributional clustering of words for text classification », *Actes de la 21e International Conference on Research and Development in Information Retrieval SIGIR'98*, 1998, p. 96-103.

Church K., Hanks P., « Word association norms, mutual information and lexicography », *Actes de la 27e Annual Conference of the Association of Computational Linguistics ACL'89,* 1989, p. 76-82.

Clavier V., Cleuziou G., Martin L., « Organisation conceptuelle de mots pour la recherche d'information sur le web », *Actes de la Conférence Francophone d'Apprentissage CAp'02*, Orléans, juin 2002, Presses Universitaires de Grenoble, p. 220-235.

Cleuziou G., Martin L ., Vrain C., « PoBOC : un algorithme de "soft-clustering". Applications à l'apprentissage de règles et au traitement de données textuelles », *Actes des 4e Journées d'Extraction et Gestion des Connaissances EGC'04*, Clermont-Ferrand, Janvier 2004, RNTI numéro spécial, p. 217-228.

Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R., « Indexing by latent semantic analysis », *Journal of the American Society for Information Science*, vol. 41, n°6, 1990, p. 391-407.

Dhillon I., Mallela S., Kumar R., « Enhanced Word Clustering for Hierarchical Text Classification », *Actes de la 8e International Conference on Knowledge Discovery and Data mining SIGKDD'02*, Canada, 2002, ACM press, p. 191-200.

Dhillon I., Mallela S., Kumar R., « A divisive Information Theoretic feature clustering algorithm for text classification », *Journal of Machine Learning Research*, vol. 3, 2003, p. 1265-1287.

Diday E., Une représentation visuelle des classes empietantes, Rapport INRIA n°291, 1984.

Fano R., *Transmission of Information : A Statistical Theory of Communication*, MIT Press and John Willey & Sons, 1961.

Greffenstette G., *Exploration in Automatic Thesaurus Discovery*, Londres, Kluwer Academic Publishers, 1994.

Jain A., Murty M., Flynn P., « Data Clustering: A Review », *ACM Computing Surveys*, vol. 31, n°3, 1999, p. 264-323.

Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., « Projet NeuroWeb : un moteur de recherche multilingue et cartographique », *Actes de la 5e Conference Hypertexts et Hypermédias H2PTM'99*, 1999, Paris.

Lewis D., Representation and learning in information retrieval, Ph.D. thesis, University of Massachusetts, 1991.

Lin K., Kondadadi R., « A word-based soft clustering algorithm for documents », *Actes de la 16e International Conference on Computers and Their Applications CATA'01*, Etats-Unis, mars 2001, p. 391-395.

Liu H., Setiono R., « Chi2: Feature selection and discretization of numeric attributes », *Actes de la 7e International Conference on Tools with Artificial Intelligence ICTAI'95*, Etats-Unis, 1995, p. 388-391.

MacQueen J., « Some methods for classification and analysis of multivariate observations », *Actes du 5e Berkeley Symposium on Mathematical statistics and probability*, vol. 1, 1967, University of California Press, p. 281-297.

Pereira F., Tishby N., Lee L., « Distributional clustering of English words », *Actes du 30e Annual Meeting of the ACL,* 1993, p. 183-190.

Porter M., « An algorithm for suffix stripping », *Program*, vol. 14, n°3, 1980, p. 130-137.

Resnik P., « Using information content to evaluate semantic similarity in a taxonomy », *Actes de la 14e International Joint Conference on Artificial Intelligence IJCAI'95*, Montreal, août 1995, p. 448-453.

Slonim N., Tishby N., « Document clustering using word clusters via the information bottleneck method », *Actes de la 23e International Conference on Research and Development in Information Retrieval SIGIR'00*, Athènes, 2000, p. 208-215.

Sneath P., Sokal R., *Numerical taxonomy. The principles and practice of numerical classification*, San Francisco, W. H. Freeman, 1973.

Turney P., « Mining the Web for synonyms: PMI-IR versus LSA on TOEFL », *Actes de la 12e European Conference on Machine Learning ECML'01*, Allemagne, 2001, p. 491-502.

Yang Y., Pedersen J., « A Comparative Study on Feature Selection in Text Categorization », *Actes de la 14e International Conference on Machine Learning ICML'97*, 1997, p. 412-420.