Une technique de réétiquetage dans un contexte de catégorisation de textes

Jérémy Clech — Djamel A. Zighed

Laboratoire E.R.I.C.
Université Lumière - Lyon2
5, ave Pierre Mendès-France
F-69676 Bron cedex
jclech@eric.univ-lyon2.fr ; zighed@univ-lyon2.fr

RÉSUMÉ. *Par essence, l'apprentissage supervisé nécessite que chaque individu de l'ensemble d'apprentissage soit préalablement étiqueté. Dans un contexte de catégorisation de textes, l'étiquetage consiste à affecter les catégories d'appartenance d'un document. Cette opération est réalisée par un expert et peut être perçue comme subjective puisque basée sur l'interprétation du document. Ainsi, l'étiquetage peut être considéré comme inconsistant dans certain cas. Par exemple, deux experts peuvent étiqueter différemment un même document, ou encore un même expert peut étiqueter différemment un même document soumis à deux instants différents. Cette inconsistance peut affecter l'efficacité du classifieur. Pour atténuer ces inconvénients, nous considérons que certains individus de l'ensemble d'apprentissage sont mal étiquetés et doivent être réétiquetés. Dans cet article, nous utilisons une méthode de relaxation afin d'optimiser la cohérence de l'étiquetage. Nous appliquons cette technique sur un corpus bien connu dans la communauté de la catégorisation de textes : la collection Reuters-21578 ApteMod. Nous montrons sur ces données que ce type de prétraitement apporte de large bénéfices.*

ABSTRACT. *In supervised learning, all the instances of the learning sample must be pre-classified. In some cases, the labelling is done subjectively by an expert. For instance, in text categorisation, each text is assigned to one or more labels or categories. In such cases, the labelling is inconsistent because it may be different from an expert to another and even may differ on the same expert at different times. The subjectivity of the labelling affects certainly the reliability of the classifier. To reduce the consequences of the subjective labelling, we consider that some examples of the learning sample are misclassified and should be restored. In this article, we use a relaxation method to optimize the coherence of the labelling. We have applied this technique on a well-known text categorization collection: the Reuters-21578 ApteMod collection. We show that this data pre-processing brings some strong benefits.*

MOTS-CLÉS : *réétiquetage, graphe de voisinage, relaxation, catégorisation de textes.*

KEYWORDS: *re-labelling, neighbourhood graph, relaxation, text categorization.*

DN – 8/2004. Fouille de textes, pages 55 à 69

1. Introduction

La catégorisation de textes consiste à identifier un lien fonctionnel ente un ensemble de catégories (ou étiquettes ou encore sujets) et un ensemble de textes. Usuellement, ce lien fonctionnel, plus communément appelé modèle prédictif, est estimé à l'aide de méthodes d'apprentissage supervisé. Un ensemble de textes préétiquetés, dénommé ensemble d'apprentissage, est nécessaire afin d'estimer les paramètres du modèle le plus efficace, c'est-à-dire celui ayant le plus faible taux d'erreur pour un nombre constant de paramètres.

L'objectif de la catégorisation de textes est d'être capable d'affecter automatiquement les étiquettes d'un ensemble de nouveaux textes. Une application typique est le classement automatique de dépêches de presse, à l'aide de leur contenu, en différentes thématiques (*i.e.* « les actualités du monde », « l'économie » et « les sports ») (Moulinier, 1996).

Plus précisément, dans la catégorisation de documents, chaque texte peut appartenir à un ensemble d'étiquettes en fonction de leur(s) thématique(s). Usuellement, l'étiquetage de l'ensemble d'apprentissage est réalisé par un expert. Ainsi, pour un texte donné, le préétiquetage peut varier en fonction de qui l'a étiqueté et à quel moment. Dans certains cas, ce préétiquetage peut être considéré comme inconsistant.

Comme nous l'avons rappelé, le processus d'apprentissage génère un modèle capable de produire la probabilité qu'un nouveau document a d'appartenir à chacune des étiquettes définies dans l'ensemble d'apprentissage. Cependant, si cet ensemble est inconsistant, le modèle peut être de mauvaise qualité. Pour diminuer ces effets, une possibilité consisterait à regrouper un comité d'experts convenant ensemble d'un unique étiquetage pour chacun des documents de l'ensemble d'apprentissage. En effet, le gain de cette opération serait la réduction certaine de l'hétérogénéité entre les experts. Cependant, et spécifiquement pour les grands corpus, une quantité d'efforts et de temps est considérable pour obtenir un étiquetage consensuel pour l'ensemble des textes.

Les textes pour lesquels les étiquettes associées seraient susceptibles de varier en fonction de l'expert ou du moment de l'étiquetage, peuvent être perçus comme des cas mal étiquetés ou comme du bruit. En outre, (Brodley *et al.*, 1996 ; 1999) ont montré que lorsque le bruit concerne seulement l'ensemble d'apprentissage alors la suppression des individus mal étiquetés améliore considérablement les résultats de l'étape de généralisation, à condition que le ratio de bruit ne dépasse pas 20 % voire 40 % dans certains cas. Ce point de vue argumente que la subjectivité de l'étiquetage pénalise l'étape de généralisation.

Par ailleurs, pour le traitement de données qualitatives mal étiquetées, (Wilson, 1972 ; John, 1995 ; Brodley *et al.*, 1996, 1999) ont exploré la détection et suppression de telles données tandis que Lallich *et al.* (2002) et Muhlenbach (2002) ont choisi la détection pour le réétiquetage.

Nous proposons ici une approche consistant à réétiqueter automatiquement l'ensemble d'apprentissage avant de réaliser le processus d'apprentissage. Cette étape peut être considérée comme étant un prétraitement de nos données, l'objectif étant de réduire le bruit sur l'ensemble d'apprentissage en atténuant les conséquences d'un étiquetage subjectif. Pour ce faire, nous utilisons un processus itératif de relaxation autorisant la modification du préétiquetage sous réserve d'améliorer l'homogénéité de ce dernier. En d'autres termes, ce processus examine toutes les étiquettes de chacun des textes de l'ensemble d'apprentissage et change certaines d'entre elles à condition d'obtenir une meilleure consistance (ce qui est équivalent à diminuer l'inconsistance). Cette inconsistance est évaluée par un critère mathématique combinant deux notions. La première est la similarité (S) évaluée à partir de la similarité entre l'étiquetage initial et l'étiquetage courant. La seconde est la cohérence locale moyenne (CLM), calculée en fonction des étiquettes du voisinage de chacun des textes. La propriété principale recherchée peut se résumer en ces mots : le réétiquetage doit d'une part rester le plus similaire possible de l'étiquetage initial, et d'autre part, les étiquettes de chaque texte doivent être similaires à celles des textes de leur voisinage.

Cet article est organisé comme suit : les notations utilisées ici sont présentées dans la section suivante. La section 3 décrit la notion de graphe de voisinage. La section 4 définit la technique de relaxation et ses deux critères. En section 5, cette technique est appliquée sur la très connue collection de Reuters-21578 (ApteMod). Enfin, les résultats obtenus sont discutés en section 6.

2. Notations

L'ensemble d'apprentissage Ω est composé de *n* textes préétiquetés composant l'étiquetage initial.

Soit *K* l'ensemble des étiquettes et soit *r* le nombre d'étiquettes pouvant être affectées à chacun des textes, $r = |K|$. Ces étiquettes forment l'ensemble des variables exogènes du problème de catégorisation de textes.

Pour un texte $i \in \Omega$ et une étiquette $k \in K$, nous définissons $p_{i,k}$ comme la probabilité d'appartenance de *i* à l'étiquette *k*. Nous associons à chaque texte *i* un vecteur $P_i = (p_{i,1}, \cdots, p_{i,r}) \in [0,1]^r$ donnant la probabilité d'appartenance de *i* pour chacune des *r* étiquettes ; l'étiquetage initial du texte *i* est noté P_i^0. Nous faisons remarquer que la somme des probabilités d'appartenance pour un texte *i* n'est pas nécessairement égale à 1 puisque nous considérons qu'un document peut appartenir à plusieurs étiquettes.

Par ailleurs, pour l'étiquetage initial la probabilité $p_{i,k}^0$ qu'un texte *i* appartienne à l'étiquette *k* est égale à 1 si l'expert juge que le document appartient effectivement à cette étiquette, 0 sinon.

Enfin, pour le corpus entier, l'étiquetage initial est noté $\mathcal{P}^0$ et $\mathcal{P}$ note l'étiquetage courant.

Pour chaque texte $i \in \Omega$, nous lui associons le vecteur X_i suivant :

$$X_i = \left(X_i^1, X_i^2, \ldots, X_i^j, \ldots, X_i^p \right) \in \mathbb{R}^p$$

où X^j est un attribut quantitatif et les X^j pour $j \in [1, \ldots, p]$ sont supposés être sélectionnés par une méthode de prétraitement appropriée aux données textuelles. Par exemple, l'ensemble d'attributs ainsi formé peut provenir de l'ensemble des mots présents dans le corpus et les X_i^j dénombrent les occurrences du mot *j* dans le texte *i*. Il existe un grand nombre de codages possibles des textes, voir (Scott *et al.*, 1999) pour plus de détails.

3. Graphe de voisinage

La première étape de notre méthode consiste à définir un graphe de voisinage afin de pouvoir établir une mesure de cohérence.

En accord avec l'ensemble des attributs quantitatifs X^j pour $j \in [1, \ldots, p]$, chaque texte peut être vu comme étant un point de l'espace $\mathbb{R}^p$. A cet espace nous associons une métrique comme par exemple la distance euclidienne ou encore la métrique cosinus.

Dès lors, nous pouvons construire un graphe de voisinage où chacun des nœuds représente un texte. Une arête entre deux nœuds indique que ces deux nœuds (textes) vérifient une relation binaire spécifique. Cette relation est symétrique et dépend du graphe de voisinage. Les deux nœuds ainsi reliés sont appelés voisins.

Considérons un corpus de textes. Pour construire un graphe de voisinage sur cet ensemble, il existe plusieurs méthodes. Nous présentons dans les sous-sections suivantes quatre méthodes (Preparata *et al.*, 1985 ; Muhlenbach, 2002).

En complément des notations de la section 2, nous notons par $\mathcal{N}_i$ l'ensemble des voisins du texte *i* défini par la méthode de graphe de voisinage choisie.

3.1. *Arbre de recouvrement minimum*

L'arbre de recouvrement minimum est un graphe de voisinage où la somme des longueurs des arêtes composant le graphe est minimale. Comme l'illustre la

figure 1, ce graphe est sans cycle, donc est un arbre et dont le nombre d'arêtes est égal à *(n-1)*.

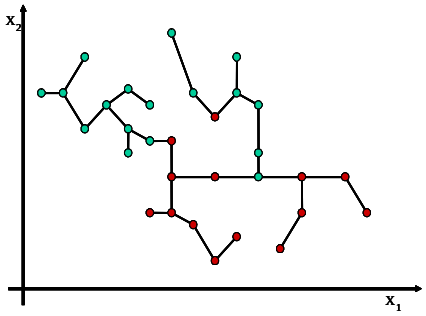

Figure 1. *Exemple d'arbre de recouvrement minimum dans* $\mathbb{R}^2$

3.2. *Polyèdres de Delaunay*

Dans $\mathbb{R}^p$, *(p+1)* points définissent un polyèdre de Delaunay si et seulement si l'hypersphère circonscrite aux *(p+1)* points ne contient aucun point. Le graphe des polyèdres de Delaunay sera défini comme étant l'union des polyèdres de Delaunay. Dans $\mathbb{R}^2$, comme sur la figure 2, on parle de triangulation de Delaunay, et les points α, β et γ forment un triangle de Delaunay car le cercle circonscrit à ces points ne contient aucun autre point.

Cette méthode détermine pour chaque point un nombre important de voisins. Cependant, sa complexité algorithmique est très élevée.

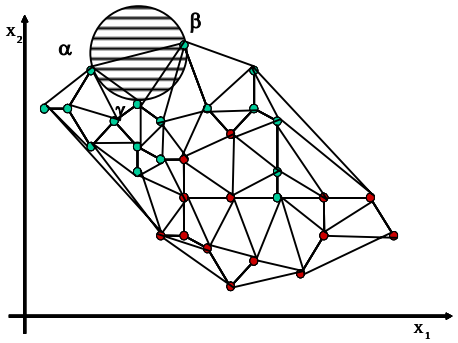

Figure 2. *Exemple de triangulation de Delaunay dans* $\mathbb{R}^2$

3.3. *Graphe de Gabriel*

Le graphe de Gabriel est un graphe où deux points α et β sont voisins si ils vérifient la propriété suivante : l'hypersphère de diamètre $[\alpha, \beta]$ doit être vide, *i.e.* qu'elle ne contient aucun autre point (voir figure 3).

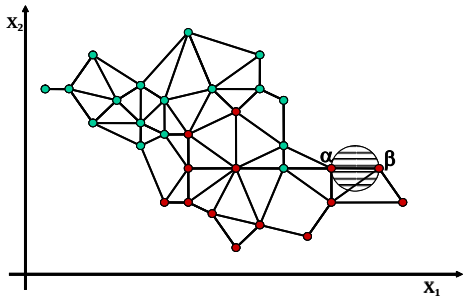

Figure 3. *Exemple de graphe de Gabriel dans* $\mathbb{R}^2$

3.4. *Graphe des voisins relatifs*

Le graphe des Voisins Relatifs est un graphe où deux points α et β sont voisins si ils vérifient la propriété suivante : la lunule, correspondant à la région hachurée dans la figure 4, doit être vide. En d'autres termes, les points α et β sont voisins si l'équation [1] est vérifiée :

$$d(\alpha,\beta) \leq Max(d(\alpha,\gamma), d(\beta,\gamma)),\ \forall \gamma \in \Omega \setminus \gamma \neq \alpha, \beta \qquad [1]$$

où $d(a,b)$ est la distance entre $a, b \in \Omega$ dans $\mathbb{R}^p$.

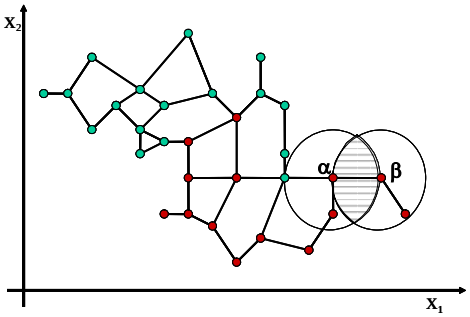

Figure 4. *Exemple de graphe des voisins relatifs dans* $\mathbb{R}^2$

4. Relaxation multi-étiquettes

Les techniques de relaxation proviennent de méthodes de traitements d'images où elles ont été utilisées à l'origine pour restaurer les niveaux de gris d'images altérées. Ces techniques de restauration s'effectuent de manière itérative et s'appuient sur la valeur des pixels voisins. Dans ce papier, nous voulons évaluer si ce genre de technique peut apporter de larges bénéfices pour le traitement particulier de la catégorisation de document textuel où chaque document appartient à un ensemble de catégories.

Dans cette section, nous définissons dans un premier temps les deux critères utilisés : la similarité et la cohérence locale moyenne. Nous détaillerons ensuite la mise en œuvre de la relaxation afin de rechercher l'optimum global.

4.1. *Similarité*

Le résultat de la relaxation est de modifier les probabilités d'appartenance à certaines étiquettes. En accord avec nos notations, $\mathcal{P}^0$ désigne l'étiquetage initial et $\mathcal{P}$ l'étiquetage courant. Le rôle de la propriété de similarité est de s'assurer que $\mathcal{P}$ ne s'éloigne pas trop de l'étiquetage initial. En effet, même si nos données peuvent être considérées comme mal étiquetées, il nous faut préserver « au mieux » l'information globale du corpus. Nous rappelons que notre objectif est de rendre plus cohérent l'étiquetage et non de l'éloigner de l'information qu'il peut contenir.

Pour mesurer la similarité par rapport à l'étiquetage initial, nous nous basons sur les écarts des probabilités d'appartenance pour chaque texte du corpus entre l'étiquetage initial et l'étiquetage courant :

$$S(\mathcal{P}^0, \mathcal{P}) = \sum_{i \in \Omega} \sum_{k=1}^{r} \left(P_{i,k}^0 - P_{i,k} \right)^2 \qquad [2]$$

4.2. *Cohérence locale moyenne*

Tous les graphes de voisinages présentés précédemment modélisent la similarité entre les textes représentés comme des points dans l'espace $\mathbb{R}^p$. Le graphe des voisins relatifs (GVR) contient le graphe de l'arbre de recouvrement minimum (ARM). D'un autre côté, il est contenu dans le graphe de Gabriel (GG) qui est lui-même contenu dans le graphe des polyèdres de Delaunay (GPD) (Preparata *et al.*, 1985). Ces relations d'inclusions sont résumées par l'équation [3].

$$ARM \subseteq GVR \subseteq GG \subseteq GPD \qquad [3]$$

L'ARM et le GPD nous sont faiblement informatifs en raison respectivement d'une trop faible quantité ou d'un trop grand nombre d'arêtes. En pratique, le GG et le GVR donnent des performances similaires. Ici, nous utilisons le graphe des voisins relatifs (Toussaint, 1980).

Dans un contexte de relaxation, nous supposons que les étiquettes d'un document se trouvent être « relativement » proches des étiquettes des documents voisins. Afin de mesurer cette dissimilarité relative, nous définissons la cohérence locale entre un document et ses voisins telle que plus leurs étiquettes sont similaires, plus leur cohérence locale tend vers 0 et plus leurs étiquettes sont différentes plus la cohérence locale tend vers *r*. Ainsi, en généralisant à l'ensemble du corpus, nous définissons la cohérence locale moyenne (CLM) comme étant la moyenne des cohérences locales de chacun des textes :

$$CLM(\mathcal{P}) = \sum_{i\in\Omega}\left[\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i}\left(\sum_{k=1}^{r}\left(P_{i,k}-P_{j,k}\right)^2\right)\right] \quad [4]$$

4.3. *Définition de l'optimum global*

En s'appuyant sur ces deux critères, le processus de relaxation modifie les probabilités d'appartenance aux étiquettes de telle façon à maximiser la cohérence globale moyenne tout en restant le plus proche possible de l'étiquetage initial. Mathématiquement, nous définissons dans l'équation [5] la fonction *F* comme la somme de la fonction de similarité *S* (cf. équation [2]) et de la fonction de cohérence locale moyenne *CLM* (cf. équation [4]). Le processus de relaxation vise à minimiser cette équation dans le but de trouver l'étiquetage optimal.

Le paramètre λ permet de déterminer la confiance envers l'étiquetage initial :

$$F(\mathcal{P}) = \lambda S(\mathcal{P},\mathcal{P}^0) + CLM(\mathcal{P})\ ;\ \lambda \in \mathbb{R}^{+*} \quad [5]$$

Pour démontrer que $F(\mathcal{P})$ possède un minimum unique, Largeron (1991) se base sur la propriété que les fonctions strictement convexes admettent un minimum unique. Elle a prouvé la convexité stricte de *F* en démontrant que *F* est la composée de deux fonctions strictement convexes (*S* et *CLM*). Ainsi, *F* admet un minimum unique et nous notons cet étiquetage optimal $\mathcal{P}^*$. Par ailleurs, cet optimum peut être estimé itérativement par l'équation [6] (Zighed *et al.*, 1990) : à partir de l'étiquetage obtenu à l'itération *m*, les probabilités d'appartenance sont modifiées en fonction des valeurs des documents voisins et convergent vers l'optimum global.

$$\forall i \in \Omega \; ; \; p_{i,k}^{m+1} = \frac{\lambda p_{i,k}^{0} + \sum_{j \in \mathcal{N}_i} \left[\left(\frac{1}{|\mathcal{N}_i|} + \frac{1}{|\mathcal{N}_j|} \right) p_{j,k}^{m_0} \right]}{1 + \lambda + \sum_{j \in \mathcal{N}_i} \frac{1}{|\mathcal{N}_j|}}, \quad [6]$$

$$\text{avec } k \in [1, \ldots, r] \text{ et } m_0 = \begin{cases} m & \text{si } j > i \\ m+1 & \text{si } j < i \end{cases};$$

L'algorithme qui implémente l'équation [6] est donné ci-dessous. Il prend en entrées deux paramètres ε et λ. Le premier permet d'indiquer le seuil à partir duquel nous considérons qu'il y a convergence et le second détermine la confiance de l'étiquetage initial. Le paramètre d'entrée-sortie correspond à l'étiquetage du corpus.

```
Programme Relaxation(entrées: epsilon,lambda;
                     entrée-sortie: P[])
Début
    Répéter
      eta := 0;
      Pour chaque i appartenant à Omega Faire
      Début
        Pour chaque k appartenant à K Faire
        Début
          numerateur := 0; denominateur := 0; tmp := 0;
          Pour chaque j appartenant au voisinage de i
          Début
             numerateur := numerateur + P[j,k] *
                           (1/card(voisinage de i) +
                           1/card(voisinage de j));
             denominateur := denominateur +
                           1/card(voisinage de j);
          Fin;
          tmp := P[i,k];
          P[i,k] := (lambda * P[i,k]+numerateur)/(1+
                    lambda + denominateur);
          tmp := |tmp – P[i,k]|;
          eta := eta + tmp;
        Fin;
      Fin;
    Jusqu'à (eta < epsilon) //Critère de convergence
Fin.
```

5. Résultats expérimentaux basés sur la collection Reuters-21578 ApteMod

Afin d'expérimenter notre technique de relaxation dans le cadre de la catégorisation de textes et afin de rendre comparables nos résultats, nous avons

choisi la collection Reuters (Apte *et al.*, 1994 ; Lewis *et al.*, 1994 ; Cohen *et al.*, 1996 ; Yang *et al.*, 1997, 1999). Dans cette section, nous décrivons succinctement la collection Reuters-21578 ApteMod, et décrivons les différents paramétrages effectués pour la réalisation de nos expérimentations. Ensuite, nous définissons les mesures de performances utilisées et présentons ensuite les résultats obtenus.

5.1. *La collection Reuters-21578 ApteMod*

Comme expliquée dans (Yang *et al.*, 1999), cette collection est une version récente de ce corpus obtenu par la suppression des documents non étiquetés comme dans Reuters-22173 (Yang *et al.*, 1997), et d'autre part par la sélection des catégories ayant au moins un document dans l'ensemble d'apprentissage et un dans l'ensemble de test. Il en résulte $r = 90$ catégories avec $n = 7\,769$ documents pour l'ensemble d'apprentissage et 3 019 pour l'ensemble de test[1]. Le vocabulaire utilisé est riche : environ 17 000 mots après désuffixation et suppression des mots-outils. Le nombre de catégories par document est hautement déséquilibré : 1,3 en moyenne avec un maximum de 14 catégories pour quelques documents. Pour plus de détails sur cette collection, voir (Yang *et al.*, 1999).

5.2. *Paramétrages*

Afin de préparer le corpus pour la catégorisation, nous appliquons les mêmes prétraitements que dans (Yang *et al.*, 1997) : les documents sont convertis en minuscule, les mots outils sont supprimés et nous avons appliqué l'algorithme de désuffixation de Porter (Porter, 1980).

Nous appliquons la méthode de sélection d'attribut du $\chi^2_{\max}$ et sélectionnons ainsi les 1 000 meilleurs termes ; en accord avec nos notations, nous avons alors $p = 1\,000$. Le choix de 1 000 est un compromis entre l'obtention des meilleurs résultats (obtenus pour 2 000 variables par (Yang *et al.*, 1997)) et l'économie de temps de calcul. La pondération des occurrences se base sur le TFxIDF (*Term Frequency Inverse Document Frequency*).

Après construction du graphe de voisinage sur l'ensemble d'apprentissage, nous exécutons le processus de relaxation. Nous avons utilisé la métrique cosinus (Rajman *et al.*, 1998) pour l'élaboration du graphe des voisins relatifs. Enfin, nous avons fixé différentes valeurs de λ en fonction de la confiance que nous associons à la qualité du préétiquetage.

Pour pouvoir mesurer l'impact de la relaxation, nous avons comparé l'évolution de l'efficacité d'un classifieur sur le même ensemble test, à partir des données brutes puis à partir des données relaxées. Comme classifieur, nous avons utilisé les

1. Le découpage de cette collection est disponible à http://www-2.cs.cmu.edu/~yiming/

k plus proches voisins (kPPV) (Mitchell, 1997) puisque il est l'un des meilleurs algorithmes pour la classification de textes et en particulier sur cette collection comme montré dans (Yang *et al.*, 1999). Nous avons fixé *k* à 10.

5.3. *Mesures de performance*

Les performances en termes de classification sont généralement mesurées à partir des traditionnelles mesures de rappel (équation [7]) et précision (équation [8]) issues du domaine de la recherche d'information (Salton, 1989) :

$$\text{rappel}=\frac{\text{catégories trouvées et correctes}}{\text{nombre de catégories correctes}} \quad [7]$$

$$\text{précision}=\frac{\text{catégories trouvées et correctes}}{\text{nombre de catégories trouvées}} \quad [8]$$

Cependant, pour chaque document, le kPPV retourne un vecteur contenant un score d'appartenance pour chaque catégorie. Il nous faut alors déterminer un seuil au-delà duquel le document est considéré comme appartenant effectivement à cette catégorie. Dans notre cas, voulant davantage mesurer l'impact de la relaxation que l'efficacité du classifieur à proprement parler, nous utilisons la méthode de la précision moyenne des 11 points suggérée par (Salton, 1989) : 11 seuils de rappels sont définis de 0 % à 100 % par pas de 10 % et on calcule la valeur de la précision pour chacune de ces valeurs. La moyenne de ces 11 valeurs de précision estime la capacité de catégoriser un document. La moyenne de ces résultats obtenus pour les différents textes de l'ensemble de test permet d'évaluer la capacité globale du classifieur sur ce corpus.

5.4. *Résultats expérimentaux*

Dans le cadre de nos expérimentations, nous avons choisi la collection Reuters-21578 ApteMod puisque cette dernière a fait l'objet d'une attention particulière quant à sa qualité. Ainsi son étiquetage initial peut être considéré comme idéal, et nous considérons le score de précision moyenne obtenu à partir de ce corpus (n'ayant subi aucune déformation de notre part), comme étant notre score de référence.

Même si cet étiquetage nous semble idéal dans le sens où il a été contrôlé par plusieurs experts, nous appliquons notre étape de relaxation dans le but d'évaluer la stabilité des résultats avant et après relaxation. Logiquement, le processus de relaxation ne devrait modifier qu'un faible nombre d'étiquettes puisque nous allons associer à λ une valeur suffisamment élevée traduisant notre confiance envers la qualité de l'étiquetage initial. D'un autre côté, afin de nous retrouver dans une

situation où l'étiquetage contient du bruit, nous allons modifier aléatoirement la valeur de 10 % des étiquettes de l'ensemble d'apprentissage. Dans ce cas, nous attendons que les résultats obtenus après relaxation se rapprochent sensiblement du score de référence.

Ainsi, nous avons effectué 4 expériences soumises aux mêmes conditions d'apprentissage et de tests, mais en utilisant 4 prétraitements différents pour l'ensemble d'apprentissage :

1. *Aucun prétraitement :* l'ensemble d'apprentissage est inchangé ;

2. *Relaxation :* nous appliquons l'étape de relaxation sur l'ensemble d'apprentissage en fixant λ à 10 en raison de notre forte confiance envers l'étiquetage initial ;

3. *Bruitage et relaxation :* nous utilisons l'ensemble d'apprentissage bruité à hauteur de 10 % auquel nous appliquons une étape de relaxation en fixant lambda à 0,25 en raison de notre faible confiance envers le préétiquetage bruité ;

4. *Bruitage :* nous utilisons l'ensemble d'apprentissage bruité.

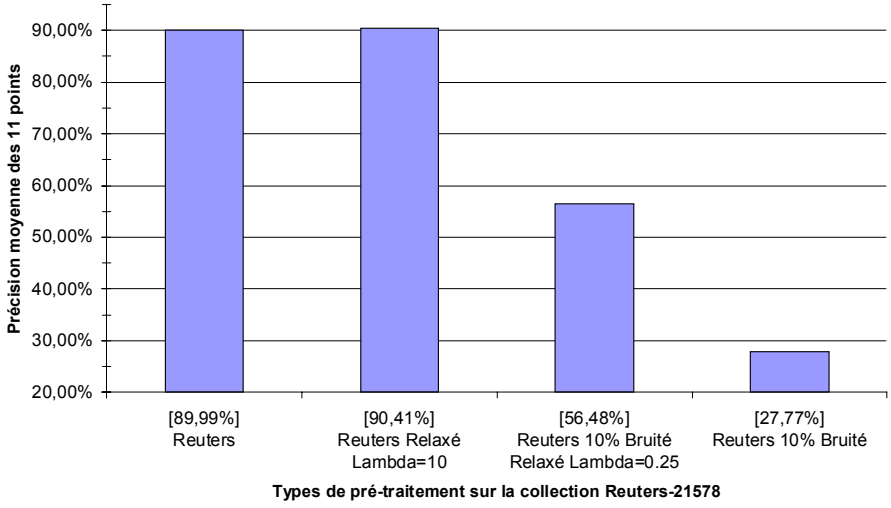

Figure 5. *Précision moyenne du kPPV en fonction des prétraitements de la collection Reuters-21578*

Evidemment, nous utilisons le même corpus bruité pour les étapes 3 et 4. Le bruitage a été effectué en inversant la valeur d'une étiquette sélectionnée aléatoirement, *i.e.* si sa valeur était 0, elle a été remplacée par 1 et réciproquement. Nous avons sélectionné aléatoirement 10 % des étiquettes d'un total de 7 769

documents par 90 étiquettes, soit un changement de 6 9921 étiquettes. En d'autres termes, nous avons changé en moyenne 9 étiquettes par texte.

La figure 4 montre la précision moyenne obtenue pour chacune des expériences effectuées. Comme illustré, la première expérience (sans prétraitement) retourne un score proche des 90 %, score se trouvant en adéquation avec d'autres expériences effectuées sous des conditions similaires comme dans (Yang *et al.*, 1999). La deuxième expérience retourne un score similaire à la première, *i.e.* 90 %. La troisième expérience renvoie un score proche de 60 %, tandis que la dernière a un score de l'ordre de 30 %, soit, comme attendu, le plus mauvais score.

6. Discussion

Ces expérimentations montrent que la relaxation peut apporter des gains substantiels. En effet, l'utilisation de la relaxation sur le corpus bruité apporte un gain spectaculaire de l'ordre de 30 %. De plus, les résultats sont similaires avant et après relaxation de la collection originale, ce qui montre l'efficacité de la similarité (cf. 4.1).

Nous en déduisons que ce type de préparation des données apporte des gains sur des collections bruitées ou non. Cependant, le type de bruit utilisé ici n'est pas réellement comparable avec la subjectivité de l'étiquetage par un expert puisque ce dernier est en principe capable d'argumenter ses choix. Ainsi, nous travaillons à tester notre méthode sur des corpus dont l'étiquetage est réputé douteux. Cependant, cela risque de provoquer des problèmes d'évaluation. En effet, en supposant que nous travaillons sur un tel corpus, comment s'assurer de la validité de l'étiquetage de l'ensemble de test, si ce n'est qu'en demandant à des experts d'évaluer les réponses ?

Pour pallier cet inconvénient, nous cherchons également à mieux modéliser les divergences d'étiquetage entre des experts. Une hypothèse raisonnable semble être que pour un texte jugé par deux experts, les étiquettes qui diffèrent sont sémantiquement proches. Des groupes d'étiquettes peuvent alors être définis formant ainsi une hiérarchie d'étiquettes. Dès lors, nous pouvons inclure du bruit sur l'étiquetage initial en modifiant une étiquette par le choix aléatoire d'une des autres étiquettes appartenant au même groupe.

L'étape de relaxation peut paraître coûteuse puisque l'étape élaborant le graphe de voisinage de l'ensemble d'apprentissage a une complexité en $O(n^3)$. Néanmoins, cette opération ne doit être effectuée qu'une seule fois puisqu'elle ne s'applique que sur l'ensemble d'apprentissage. Et il est certainement moins coûteux que d'exiger un étiquetage manuel de grande qualité.

Enfin, nous avons effectué ces mêmes expériences en utilisant divers types de graphes de voisinage mais au final les différences étant infimes nous ne les présentons pas.

7. Conclusions et perspectives

Nous avons décrit dans ce papier une méthode de relaxation permettant de rendre plus consistant l'étiquetage initial de l'ensemble d'apprentissage d'un corpus de textes en modifiant automatiquement cet étiquetage. Nous avons en outre montré que la relaxation alliée à l'analyse de contenu des textes apporte des bénéfices évidents dans le cadre de la catégorisation de documents. Ainsi, cette méthode permet d'utiliser des corpus ayant un étiquetage douteux.

De plus, cette technique peut être appliquée dans des domaines où l'étiquetage est hautement subjectif. Cependant, nous la considérons non applicable dans les domaines où une observation est objective. Par exemple, considérons le domaine de recherche d'analyse de survie, appliquer ce type d'outils reviendrait à considérer qu'un cas constaté mort peut être considéré vivant et inversement.

Nos futurs travaux vont consister à appliquer cette méthode sur des corpus réellement bruités et de mettre en œuvre des méthodes d'évaluation des gains ainsi apportés. Enfin, il nous semble également intéressant d'étudier comment utiliser ce type d'outils afin d'optimiser directement les fonctions de coûts des classifieurs.

8. Bibliographie

Apte C., Damerau F. J., Weiss, S. M., « Towards Language-Independent Automated Learning of Text Categorization Models », *Proceedings of the 17th Annual ACM SIGIR Conference*, Dublin, IE, Springer Verlag, Heidelberg, DE, 1994, p. 23-30.

Brodley C. E., Friedl M. A., « Identifying and Eliminating Mislabeled Training Instances », *Proceedings of the 13th National Conference on Artificial Intelligence*, Portland, AAI Press, 1996, p. 799-805.

Brodley C. E., Friedl M. A., « Identifying mislabeled training data », *Journal of Artificial Intelligence Research*, vol. 11, 1999, p. 131-167.

Cohen W. W., Singer Y., « Context-sensitive learning methods for text categorization », *Proceedings of the 19th Annual ACM SIGIR Conference*, 1996, p. 307-315.

John G. H., « Robust decision trees: removing outliers from data », *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, Montréal, Québec, AAI Press, 1995, p. 174-179.

Lallich S., Muhlenbach F., Zighed, D. A., « Improving classification by removing or relabeling mislabeled instances », *Proceedings of the 13th International Symposium on Methodologies for Intelligent Systems, Foundations of Intelligent Systems*, Lyon, France, Springer-Verlag, 2002, p. 5-15.

Largeron C., Reconnaissance des formes par relaxation : un modèle d'aide à la décision, Thèse de doctorat, Université Lyon1, 1991.

Lewis D. D., Ringuette M., « Comparison of two learning algorithms for text categorization », *Proceedings of the 3rdAnnual Symposium on Document Analysis and Information Retrieval*, 1994, p. 81-93.

Mitchell T. M., *Machine Learning*, New York, McGraw-Hill, 1997.

Moulinier I., Une approche de la catégorisation de textes par l'apprentissage symbolique, Université Paris 6, 1996.

Muhlenbach F., Evaluation de la qualité de la représentation en fouille de données, Thèse de doctorat, Laboratoire ERIC, Université Lumière Lyon 2, 2002.

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

Preparata F., Shamos M., *Computational Geometry An Introduction*, New-York, Springer, 1985.

Rajman M., Lebart L., "Similarités pour données textuelles", *4th International Conference on Statistical Analysis of Textual Data (JADT'98)*, Nice, France, 1998, p. 545-555.

Salton G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Pennsylvania, Addison-Wesley, 1989.

Scott S., Matwin S., « Feature Engineering for Text Classification », *Proceedings of the 16th International Conference on Machine Learning*, San Francisco, 1999, p. 379-388.

Sebastiani F., « Machine learning in automated text categorization », *ACM Computing Surveys*, vol. 340, n°1, 2002, p. 1-47.

Toussaint G., « The relative neighborhood graph of a finite planar set », *Pattern Recognition*, vol. 12, 1980, p. 261-268.

Wilson D., « Asymptotic properties of nearest neighbors rules using edited data », *IEEE Transactions on Systems, Man and Cybernetivs*, vol. 2, 1972, p. 408-421.

Yang, Y., Liu, X., « A re-examination of text categorization methods », *Proceedings of the 22nd Annual ACM SIGIR Conference*, 1999, p. 42-49.

Yang, Y., Pedersen J. O., « A comparative Study on Feature Selection in Text Categorization », *Proceedings of the 14th International Conference on Machine Learning*, 1997, p. 412-420.

Zighed D. A., Tounissoux D., Auray J. P., Largeron C., « Discrimination basée sur un critère d'homogénéité locale », *T.S.I.*, vol. 3, 1990, p. 213-220.