# Aide à l'expertise des brevets par alignement avec les publications scientifiques

Kafil Hajlaoui, Pascal Cuxac, Jean-Charles Lamirel, Claire François

# Aide à l'expertise des brevets par alignement avec les publications scientifiques

**Kafil Hajlaoui[1], Pascal Cuxac[1], Jean Charles Lamirel[2], Claire François[1]**

*1. INIST CNRS*
*2 allée du Parc de Brabois, F-54519, Vandœuvre-lès-Nancy, France*
*prenom.nom@inist.fr*

*2. LORIA team SYNALP*
*Campus Scientifique, F-54506, Vandœuvre-lès-Nancy, France*
*jean-charles.lamirel@loria.fr*

RESUME. *Ce travail s'inscrit dans le cadre du programme de recherche QUAERO[1] , un vaste projet de recherche et d'innovation se rapportant au traitement automatique de contenus multimédias et multilingues. L'objectif abordé dans cet article est de proposer une méthode de classification automatique d'articles dans un plan de classement international de brevets relevant du même domaine. La finalité applicative de ce travail est de proposer une aide aux experts dans le processus d'évaluation de l'originalité et de la nouveauté d'un brevet, en lui proposant les citations scientifiques les plus pertinentes. Ce problème soulève de nouveaux défis en catégorisation liés du fait que le plan de classement des brevets n'est pas directement adapté à la structure des documents scientifiques et que la répartition des exemples disponibles n'est pas nécessairement équilibrée entre les différentes classes d'apprentissage. Nous proposons pour les résoudre d'appliquer une amélioration de l'algorithme des K-plus-proches-voisins (K-PPV) se basant sur l'exploitation des règles d'associations entre les termes descripteurs des documents et ceux des classes de brevets. En utilisant conjointement comme référentiels une base de brevets du domaine de la pharmacologie et une base bibliographique du même domaine issue de la collection Medline, nous montrons que cette nouvelle technique de catégorisation, qui combine les avantages des approches numériques et ceux des approches symboliques, permet d'améliorer sensiblement les performances de catégorisation, relativement aux méthodes de catégorisation usuelles, dans le cas du problème posé.*

ABSTRACT. *This paper focuses on a subtask of the QUAERO[1] research program, a major innovating research project related to the automatic processing of multimedia and multilingual content. The objective discussed in this article is to propose a new method for the*

[1] http://www.quaero.org

*classification of scientific papers, developed in the context of an international classification plan of patents related to the same field. The practical purpose of this work is to provide an assistance tool to experts in their task of evaluation of the originality and novelty of a patent, by offering to the latter the most relevant scientific citations. This issue raises new challenges in categorization research as the patent classification plan is not directly adapted to the structure of scientific documents and that there is not always a balanced distribution of the available examples within the different learning classes. We propose, as a solution to this problem, to apply an improved K-nearest-neighbors (KNN) algorithm based on the exploitation of association rules occurring between the index terms of the documents and the ones of the patent classes. By using a reference dataset of patents belonging to the field of pharmacology, on the one hand, and a bibliographic dataset of the same field issued from the Medline collection, on the other hand, we show that this new approach, which combines the advantages of both numerical and symbolical approaches, improves considerably categorization performance, as compared to the usual categorization methods.*

*MOTS-CLES : Classification supervisée, Veille scientifique et technique, Brevets, K-PPV, Règles d'association.*

*KEYWORDS: Supervised classification, Technological and scientific survey, Patents, KNN, Association rules.*

DOI:10.3199/DN.15.1-n 

## 1. Introduction

La catégorisation automatique de textes (CAT) vise à regrouper, souvent selon des thèmes communs, les documents ayant des caractéristiques spécifiques et homogènes (Cohen et Hersh, 2005). Cette approche est dite supervisée car les thèmes sont identifiés à priori, en général à l'aide d'exemples. Si ce n'est pas le cas, l'on parle alors d'approche non supervisée, de classification, ou encore, de clustering de textes. Dans ces deux types d'approche, une première étape est la transformation des documents en une représentation appropriée pour le classifieur. Cette transformation vise à pondérer et à réduire l'espace de représentation des documents, tout en ménageant la possibilité de discriminer entre ces derniers. Elle comprend usuellement des opérations de suppression des mots vides, de lemmatisation, de sélection et de pondération des descripteurs.

Dans le cas de la catégorisation, la deuxième étape est l'apprentissage : le système apprend à classer les documents selon un modèle de classement où les classes sont prédéterminées et les exemples sont connus et correctement étiquetés d'avance. Des mesures de bases simples permettent ensuite d'évaluer les résultats sur un corpus de test dont les documents sont classés en mettant en correspondance leur représentation avec celles des classes apprises. Ce sont les mesures de Rappel, également appelé Sensibilité, et de Précision. Pour une classe donnée, ces mesures sont fondées sur le nombre de documents correctement classés dans la classe, ou vrais positifs (*VP*), le nombre de documents incorrectement classés dans la classe, ou faux positifs (*FP*), et enfin, le nombre de documents de la classe qui sont classés dans une autre classe, ou faux négatifs (*FN*). Ces mesures sont généralement

moyennées sur l'ensemble des classes. La Précision moyenne mesure la capacité d'un classifieur à éviter le bruit, et le Rappel moyen, sa capacité à éviter le silence. Le comportement de ces mesures est donc généralement antagoniste : plus l'on cherche à augmenter le Rappel moyen du classifieur, plus l'on aura tendance à diminuer sa Précision moyenne, et inversement. Le F-score est donc une mesure de compromis qui représente la moyenne harmonique entre le Rappel et la Précision moyens. Ces 3 mesures s'expriment finalement comme suit :

Rappel: $R = VP/(VP + FN)$

Précision: $R = VP/(VP + FP)$

F-score: $F = 2 * (P * R)/(P + R)$

La catégorisation automatique de textes a été l'un des domaines les plus étudiés en apprentissage automatique (Hillard et al., 2007). En conséquence, une grande variété d'algorithmes de classification ont été développés et/ou évalués, souvent dans des applications telles que le filtrage des mails (Cormack, 2007) ou l'analyse des opinions et des sentiments (Pang et Lee, 2008). Dans le domaine des sciences sociales, l'apprentissage automatique a été utilisé dans la classification d'actualités (Purpura et Hillard, 2006) (Evans et al., 2007), ou des blogues (Durant et Smith, 2007). Parmi les méthodes d'apprentissage les plus souvent utilisées, figurent les réseaux de neurones (Wiener et al., 1995) (Schütze et al., 1995), les K-plus-proches-voisins (K-PPV) (Yang et Chute, 1994), les arbres de décision (Lewis et Ringuette, 1994) (Quinlan, 1986) (Apte et al., 1998), les réseaux bayésiens (Lewis, 1992) (Joachims, 1997), les machines à vecteurs supports (SVM) (Joachims, 1998), et plus récemment, les méthodes basées sur le boosting (Schapire, 1998) (Iyer et al., 2000). Bien que beaucoup de méthodes développées dans le domaine de la catégorisation automatique de textes aient permis d'atteindre des niveaux de précision appréciables lorsqu'il s'agit de textes à structure simple (par ex. courriels, résumés, etc.), il reste néanmoins encore de nombreux défis à relever dans le cas de documents complexes, ou, comme le cas que nous traitons, si le plan de classement des documents n'est pas directement adapté à leur contenu et si la répartition des exemples entre les différentes classes d'apprentissage n'est pas équilibrée.

Plusieurs travaux ont été réalisés plus spécifiquement sur des données issues de la base Medline. Ces travaux illustrent plus particulièrement l'importance des étapes de prétraitement et de représentation des données dans le cadre de la catégorisation des textes. Dans (Lan, 2007), les auteurs montrent qu'avec une représentation de textes basée sur l'approche dite « sac de mots », la pondération des termes extraits augmente significativement la performance du classifieur. Pour classer un article scientifique dans un sujet (thème), Suomela et Andrade (2005), se basent quant à eux sur la fréquence des termes, en restreignant ces derniers à des classes lexicales prédéfinies (Noms, Adjectifs, Verbes). Les auteurs évaluent leur proposition en utilisant des thèmes issus de la base Medline, et obtiennent une performance F-score de 65%. La même approche est reprise par le système MedlineRanker web-service (Fontaine et al., 2009) qui permet de retrouver une liste pertinente de notices

Medline à partir d'un ensemble de mots-clés définis par l'utilisateur. Les travaux de Yin et al. (2010) portent sur l'identification et l'extraction des interactions entre protéines à partir des articles Medline. Les documents sont traités en utilisant des bi-grammes. Avec la méthode SVM, les auteurs obtiennent une performance de 50% de vrais positifs (équiv. précision), et un taux de rappel de 51%. Récemment, la tâche d'évaluation Bio-creative III a proposé comme challenge la classification d'articles Medline spécifiques au domaine biomédical (Krallinger et al., 2010). Sur cette collection, les meilleures performances ont été obtenues, avec une précision de 89,2% et un F-score de 61,3%.

Des travaux récents s'intéressent spécifiquement à la classification des brevets, comme les travaux de (Koster et al., 2001) et de (Koster et al., 2010). Dans ces travaux, ces auteurs utilisent l'algorithme d'apprentissage heuristique de Winnow (Grow et al., 2001) et supposent pouvoir opérer sur des classes de brevets différenciées de taille homogène. Dans leurs derniers travaux, ils mènent alors deux types d'expérience : une mono-classification, où chaque brevet est classifié dans une seule classe, et une multi-classification, où chaque brevet peut être affecté à plusieurs classes. Les auteurs observent que les résultats obtenus sur le texte complet des brevets sont légèrement meilleurs que ceux obtenus sur le résumé. De même, les résultats obtenus en mono-classification semblent supérieurs à ceux obtenus en multi-classification, mais ce dernier cas reste plus difficile à juger car les référentiels utilisés par les auteurs ne sont pas identiques. Les meilleurs résultats obtenus en termes de F-score sont de 98%. Cependant, le cas de classes homogènes et dissimilaires est un cas idéal comme nous le montrons par la suite dans un contexte plus large.

L'évaluation des brevets représente, quant à elle, une opération jusqu'ici manuelle qui fait intervenir des groupes d'experts ayant des compétences dans le domaine d'analyse et qui connaissent parfaitement l'objet des brevets. Elle s'appuie sur des références et des citations vers des documents scientifiques pertinents (articles, thèses, ouvrages…). Un classement automatisé des publications dans les classes de brevets peut donc constituer une aide précieuse pour les experts. Cette démarche implique de classer des articles scientifiques (notices) dans un plan de classement des brevets ; il ne s'agit donc pas d'une problématique traditionnelle de classification automatique, comme celle présentée dans les travaux de (Koster et al., 2001) et de (Cornelis et al., 2010) décrits précédemment, car le plan de la classification utilisé n'est pas a priori adapté à une classification de notices bibliographiques d'articles scientifiques.

Dans ce nouveau contexte, deux alternatives sont possibles. Une première est de concevoir une passerelle entre le plan de classement des publications et celui des brevets. Cette démarche est cependant difficile à mettre en œuvre car elle implique l'exploitation intensive de techniques très lourdes de comparaison d'arbres (matérialisés ici par les plans de classement), et doit s'opérer en partie de manière supervisée. Une deuxième alternative est d'élaborer un système de classification des notices bibliographiques dans le plan des brevets. Elle est basée sur l'hypothèse que les citations scientifiques qui apparaissent dans un brevet sont fortement liées au domaine du brevet, donc au code de classement de ce dernier. Dans ce cadre, le

corpus d'apprentissage d'une classe donnée représentera alors l'ensemble des citations extraites des brevets de cette classe. Même si cette idée est plus facile à mettre en œuvre, elle implique néanmoins de résoudre un problème supplémentaire qui est celui d'avoir à disposition un nombre équivalent d'exemple d'apprentissage (i.e. de citations de publications) dans chacune des classes de brevets, ces classes n'ayant pas nécessairement elles-mêmes un effectif homogène, en termes de brevets.

Dans les sections suivantes, nous menons une expérimentation complète de catégorisation des publications à partir d'un corpus de brevets issus du domaine de la pharmacologie et d'un corpus bibliographique issu de la collection Medline. Dans la première section, nous présentons notre stratégie de constitution du corpus expérimental et nous illustrons les phénomènes de déséquilibre des exemples d'apprentissage et de similarité des classes qu'il est possible d'observer. En exploitant les méthodes de catégorisation usuelles, nous illustrons ensuite, dans la seconde section, l'influence de la stratégie de choix des termes descripteurs utilisés pour les documents-exemples sur les résultats de catégorisation. Deux approches sont plus particulièrement abordées, la première basée sur l'exploitation directe des mots-clés Medline, la seconde basée sur l'extraction d'index à partir du texte plein des titres et résumés en utilisant une plate-forme de traitement linguistique. Dans la troisième section, nous présentons une adaptation de l'algorithme des K-plus-proches-voisins (K-PPV) se basant sur l'exploitation des règles d'associations identifiées entre les termes descripteurs des documents et ceux des classes de brevets. Nous montrons qu'elle permet d'améliorer les résultats obtenus dans notre contexte d'apprentissage. Dans l'avant-dernière section, nous présentons de nouvelles alternatives possibles. La dernière section présente finalement notre conclusion et nos perspectives.

## 2. Constitution et indexation du corpus

### *2.1. Extraction des données*

La ressource principale de notre corpus est une collection de brevets du domaine de la pharmacologie auxquels sont associés des citations bibliographiques. La répartition des brevets dans les classes du domaine suit le code de la classification CIB (Classification Internationale des Brevets) qui est un système hiérarchique de symboles indépendants de la langue pour le classement propres à chaque domaine technologique. Le but de la CIB est de faciliter la recherche d'information en matière de brevets et d'aider à recenser les technologies existantes et nouvelles. La CIB se base, dans la tâche de la classification, sur un système de codage hiérarchique qui comporte plus de 70 000 catégories, avec des couches de plus en plus détaillées. Un exemple extrait de ce système de codage est le suivant :

A - NECESSITES COURANTES DE LA VIE
A-61- SCIENCES MEDICALE OU VETERINAIRE ; HYGIENNE
A-61-K- PREPARATIONS A USAGE MEDICAL, DENTAIRE OU POUR LA TOILETTE
A-61-K-6- PREPARATIONS POUR LA TECHNIQUE DENTAIRE ….

Dans notre cas, les notices brevet au format XML sont au nombre de 6387 réparties dans les 15 classes de la catégorie A61K (préparations à usage médical…). Comme l'illustre la figure 1, nous commençons par l'extraction des références à partir des brevets. Grace à des robots web, nous interrogeons une base de données de publications pour extraire les notices relatives aux références collectées. Chaque notice est ensuite étiquetée par la classe du brevet citant. L'ensemble des notices étiquetées représente finalement notre corpus d'apprentissage.

Nous avons interrogé la base de données Medline[2] qui est spécialisée dans le domaine de la médecine et qui, d'autre part, bénéficie de mises à jour régulières. A partir des 6387 brevets, nous avons extrait 25887 références de types bases de données, livres, encyclopédie… et articles scientifiques. L'interrogation de Medline avec les références de types articles scientifiques nous a fourni 7501 notices, ce qui représente un rappel de 90% relatif à ce type de références.

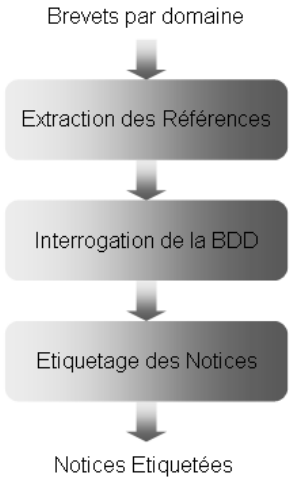


*Figure 1. Processus de construction du corpus d'apprentissage*

La figure 2 résume la répartition des notices du corpus. Au vu de la forte irrégularité de cette répartition, il semble clair qu'un des critères importants pour le choix de la méthode de classification sera sans aucun doute sa capacité à traiter le déséquilibre des exemples entre les classes. En effet, la distribution des notices entre les classes s'avère être très hétérogène : certaines des sous-classes ne contiennent que quelques dizaines de notices en comparaison alors que d'autres en contiennent plus de 2500.

[2] http://www.ncbi.nlm.nih.gov/pubmed/

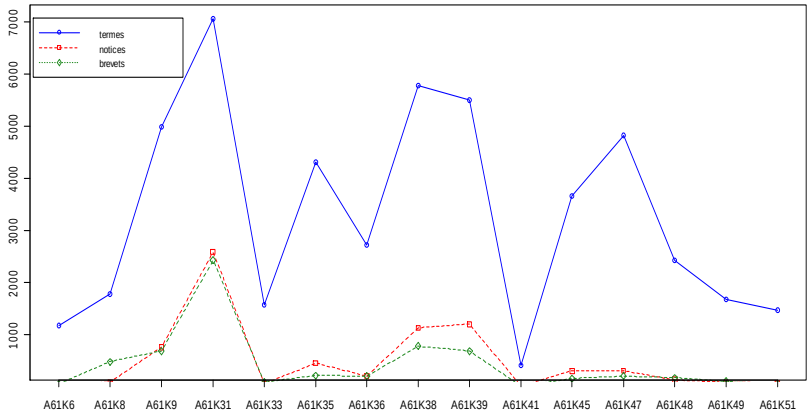


*Figure 2. Répartition des données dans les classes : brevets-notices-termes*

### *2.2. Représentation des données*

Dans la classification automatique de textes, le choix de représentation des documents est une étape cruciale. Une approche fréquente consiste à faire appel à une représentation dite « sac de mots », où la seule information utilisée est la présence et/ou la fréquence de certains mots. Dans notre contexte, nous utilisons une représentation vectorielle des documents selon le modèle de Salton (Salton, 1971), de manière à pouvoir exploiter des pondérations sur les mots. Chaque notice de la collection est ainsi représentée comme un vecteur dans un espace à $N$ dimensions, où $N$ est le nombre total des termes extraits de la collection de notices. L'ensemble de la collection des notices est représenté par une matrice de dimension $(N+1) \times J$, où $J$ est le nombre de notices dans la collection. Chaque ligne $j$ de cette matrice est un vecteur à $N$ dimensions auquel on ajoute l'étiquette de la classe pour la notice $j$. Si un descripteur $i$ n'est pas produit par la notice $j$, alors la valeur $a_{ij}$ de la matrice vaut 0. Dans le cas contraire, $a_{ij}$ prend une valeur positive. La méthode pour calculer cette valeur dépend du choix de pondération des descripteurs.

Nous avons ensuite extrait les descriptions proprement dites des notices à partir de deux approches différentes, une première basée sur les mots-clés présents dans ces dernières, et une approche alternative, basée sur les lemmes issues du traitement du texte plein des résumés à partir de méthodes de traitement automatique des langues (TAL). L'objectif de la tâche TAL dans le processus de classification automatique est d'obtenir une représentation à la fois plausible du contenu des documents et suffisamment synthétique pour être adaptée à ce processus. Cette tâche se base donc principalement sur la sélection de fragments pertinents. Dans notre cas, cette sélection s'opère à deux niveaux:

1) Tout d'abord, par la lemmatisation qui permet de diminuer fortement le nombre de schémas linguistiques en éliminant toutes les flexions et les dérivations grammaticales et,

2) D'autre part, par l'étiquetage qui consiste à assigner une catégorie grammaticale à chaque mot et qui permet par la suite de ne conserver que les catégories grammaticales jugées les plus pertinentes.

Pour ce faire nous utilisons le programme TreeTagger (Schmid, 1994) qui est à la fois un étiqueteur et un lemmatiseur développé par l'Institute for Computational Linguistics de l'Université de Stuttgart. Dans un premier temps, les documents sont lemmatisés. La suite des analyses est effectuée sur les formes lemmatisées, sauf lorsque le mot est inconnu du tagger et, dans ce cas, sa forme originale est conservée. Les signes de ponctuation et les nombres, identifiés par le tagger, sont supprimés. Un exemple plus précis de sortie du programme TreeTagger[3] est donné à la figure 3.

| | | |
|---|---|---|
| The | DT | the |
| most | RBS | most |
| widely | RB | widely |
| used | VVN | use |
| therapeutic | JJ | therapeutic |
| modality | NN | modality |
| is | VBZ | be |
| chemical | JJ | chemical |
| **pleurodesis** | **NN** | **<unknown>** |

*Figure 3. Exemple d'une phrase étiquetée et lemmatisée par TreeTagger*

La sélection d'attributs selon les catégories grammaticales permet, par exemple, d'identifier des traits de jugement subjectif pour la classification de documents par genre ou par opinion. Il est donc pertinent, dans notre cas, de mesurer l'impact de l'utilisation de descripteurs fondés sur la sélection de catégorie(s) grammaticale(s). Cette étude peut permettre de réduire de manière conséquente la taille de l'espace de description. Dans notre cas, nous avons également décidé de retenir les mots lemmatisés '<unknown>' par TreeTagger et catégorisés comme noms sous leur forme non lemmatisée (NN) car ces noms sont susceptibles de matérialiser des concepts à la fois importants et nouveaux du domaine (dans l'exemple présenté à la figure 3, c'est le cas du mot 'pleurodesis').

La pondération fréquentielle se fonde sur le nombre d'occurrences des descripteurs dans un document. Cependant, en procédant de la sorte, on donne une importance trop grande aux descripteurs qui apparaissent très souvent dans un grand nombre de documents et qui sont peu représentatifs d'un document en particulier. On trouve dans la littérature (Vincarelli, 2006) (Salton et Buckley, 1988) (Roberston et Spark Jones, 1972) une autre mesure de poids, très répandue, connue sous le nom de TF.IDF (Term Frequency. Inverse Document Frequency) qui permet de prendre en compte ce phénomène. Elle mesure l'importance d'un mot en fonction de sa

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.pdf

fréquence dans le document (TF=Term Frequency) pondérée par sa fréquence d'apparition dans tout le corpus (IDF=Inverse Document Frequency) :

$$Tf.Idf(t_k, D_j) = TF(t_k, D_j) \times Idf(t_k)$$

où $TF(t_k, D_j)$ est le nombre d'occurrences de $t_k$ dans $D_j$ et :

$$Idf(t_k) = \frac{log|S|}{DF(t\_k)}$$

où $|S|$ est le nombre de documents dans le corpus et $DF(t_k)$ est le nombre de documents contenant $t_k$.

Cette dernière mesure permet de donner un poids plus important aux mots discriminants d'un document. Inversement, un terme apparaissant dans tous les documents du corpus aura un poids faible, voire nul.

Dans l'ensemble des tests, nous appliquons deux techniques de pondération différentes selon les descripteurs extraits. Pour les lemmes, nous pondérons conjointement par la technique fréquentielle (TF) normalisée par la valeur maximum de fréquence et par la technique IDF. Pour les mots-clés, la technique TF n'a pas de sens puisque les termes d'indexation documentalistes ne sont pas redondants. C'est pourquoi nous n'utilisons dans ce cas que la technique IDF pour la pondération.

## 3. Classification

Pour évaluer la pertinence des différentes méthodes d'indexation et de pondération, nous avons choisi d'utiliser trois classifieurs pour la classification supervisée : un classifieur de type K-plus-proches-voisins (K-PPV) exploitant une distance euclidienne, un classifieur fondé sur les machines à vecteurs supports (SVM) et un classifieur probabiliste (Bayésien naïf ou BN). Le choix s'est fixé sur ces trois méthodes parce qu'il s'agit des algorithmes d'apprentissage supervisé qui donnent le plus souvent les meilleurs résultats pour la classification des textes (Sebastiani, 1999). Ces algorithmes sont utilisables sous l'environnement Weka[4].

Dans le cas de l'indexation basée sur les lemmes, nous présentons les différentes expérimentations que nous avons réalisées, en faisant varier les catégories grammaticales prises en compte dans l'indexation (A : Adjectif, N : Nom, NA : Nom+Adjectif, NV : Nom+Verbe, VA : Verbe+Adjectif, NVA : Nom+Verbe+Adjectif). Les résultats de la classification sont présentés en termes de précision et de rappel. Une précision de 100% signifie que toutes les notices sont classées dans la bonne catégorie. Cette mesure est calculée après l'application d'une validation croisée en dix sous-ensembles (pour chacun d'entre eux, 90% du corpus est utilisé pour l'apprentissage et 10% pour le test). Le rappel est le pourcentage de réponses correctes qui sont retrouvées.

[4] http://www.cs.waikato.ac.nz/ml/weka/index.html

*Tableau 1 : Résultat de la classification basée sur l'indexation par mots-clés*

| Algorithmes | K-PPV | | | | BN | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pondération | Booléen | | IDF | | Booléen | | IDF | | Booléen | | IDF | |
| | P | R | P | R | P | R | P | R | P | R | P | R |
| Mots-clés | *0,39* | *0,39* | *0,39* | 0,43 | 0,4 | **0,47** | **0,43** | 0,44 | 0,4 | 0,45 | 0,4 | 0,45 |

*Tableau 2 : Résultat de la classification basée sur l'indexation par lemmes*

| Algorithmes | K-PPV | | | | BN | | | | SVM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pondération | Fréquentiel | | TF-IDF | | Fréquentiel | | TF-IDF | | Fréquentiel | | TF-IDF | |
| | P | R | P | R | P | R | P | R | P | R | P | R |
| A | 0,42 | *0,36* | 0,42 | 0,36 | 0,38 | 0,2 | 0,37 | *0,18* | 0,45 | 0,46 | 0,45 | 0,46 |
| N | 0,5 | 0,41 | 0,52 | 0,4 | 0,43 | 0,31 | 0,44 | 0,28 | 0,54 | 0,55 | 0,54 | 0,55 |
| NA | 0,55 | 0,4 | 0,57 | 0,39 | 0,45 | 0,36 | 0,46 | 0,36 | 0,55 | 0,55 | 0,55 | 0,55 |
| NV | 0,49 | 0,38 | 0,52 | 0,38 | 0,44 | 0,35 | 0,44 | 0,31 | 0,53 | 0,54 | 0,53 | 0,54 |
| NVA | 0,6 | 0,54 | **0,61** | **0,55** | 0,44 | 0,34 | 0,45 | 0,34 | 0,54 | 0,55 | **0,55** | **0,55** |

Les tableaux 1 et 2 donnent la précision et le rappel obtenus pour la classification avec les trois algorithmes d'apprentissage sur le même corpus de notices bibliographiques, en faisant varier les méthodes d'indexation. Ils permettent de montrer à la fois que les approches utilisées pour la classification, les méthodes d'indexation et les méthodes de pondération des descripteurs ne sont pas équivalentes dans le cas du problème posé. Ainsi, les meilleurs résultats sur notre corpus sont obtenus avec la méthode K-PPV, combinée à une indexation basée sur les lemmes, impliquant les trois catégories grammaticales (Noms, Verbes, Adjectifs), et une pondération de type TF-IDF. Cette combinaison permet d'obtenir une précision de 61% et un rappel de 55%.

Ces résultats peuvent cependant être considérés comme très moyens. Ceci peut s'expliquer par le fait que les exemples d'apprentissage ne sont pas équitablement répartis entre les classes (figure 2), mais également que les classes sont très proches les unes des autres. Une similarité classe/classe a été calculée, et elle montre bien cette proximité, rendant difficile pour tout modèle la détection exacte de la bonne classe. La figure 4 montre ainsi que plus de 70% des couples de classes ont une similarité entre 0,5 et 0,9.

Nous proposons donc à la section suivante une amélioration basée sur la meilleure méthode, à savoir celle des K-PPV, et susceptible de prendre en compte les caractéristiques spécifiques du corpus, à savoir le déséquilibre entre les classes d'apprentissage et la forte similarité entre ces dernières.

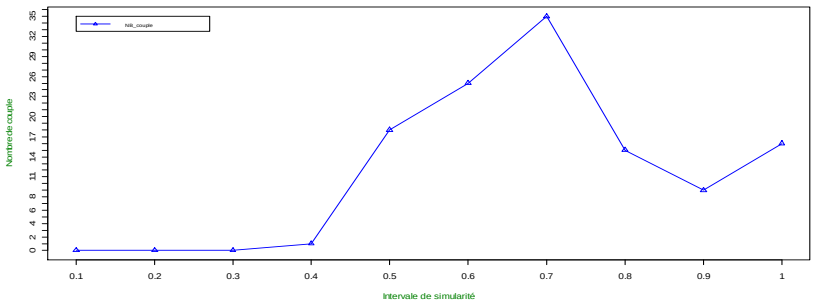


*Figure 4 : Similarité Classe/Classe*

## 4. La méthode K-PPVBA-2T

Dans cette partie, nous nous intéressons à une amélioration de l'algorithme K-PPV exploitable dans le contexte de notre problème. Nous allons présenter une définition générale des règles d'association. Ensuite, nous suggérerons une nouvelle approche pour calculer le poids des attributs des classes par l'utilisation d'un type particulier de règles d'association. Enfin, pour obtenir plus de précision dans la classification des données, nous présenterons un nouvel algorithme appelé « K-PPVBA-2T » inspiré de la méthode développée antérieurement par Mordian et al. (Mordian et al., 2009).

### *4.1. Règles d'association*

La méthode d'extraction de règles d'associations représente une méthode permettant de découvrir des relations pertinentes entre deux ou plusieurs variables. Cette méthode se base sur des lois locales et ne nécessite pas d'intervention de l'utilisateur (on laisse le système s'auto-organiser). Elle permet d'identifier, à partir d'un ensemble de transactions, un ensemble de règles qui expriment une possibilité d'association entre différents items (mots, attributs, concepts). Une transaction est une succession d'items exprimés selon un ordre donné ; de plus, des transactions différentes peuvent être de longueurs différentes.

La pertinence d'une règle d'association ainsi extraite est mesurée par son indice de support et son indice de confiance. Si on a une règle d'association : alors les indices de support et confiance sont définies par les deux équations suivantes :

$$Support = P(X \cup Y)\ ,\ Confiance = P(X|Y)$$

où $P(X \cup Y)$ indique la probabilité qu'une transaction contienne à la fois $X$ et $Y$, et $P(X|Y)$ est la probabilité conditionnelle d'avoir $Y$ sachant qu'on a $X$.

La première méthode efficace d'extraction de règles d'association a été introduite par Agrawal pour l'analyse du panier de la ménagère, par l'intermédiaire

de l'algorithme Apriori (Agrawal et Srikant, 1994). Le fonctionnement de cet algorithme peut être décomposé en deux phases :

1) Recherche des tous les « patrons » ou itemsets fréquents, qui apparaissent dans la base de données avec une fréquence supérieure ou égale à un seuil défini par l'utilisateur, appelé Min_Sup.

2) Génération, à partir de ces patrons fréquents, de l'ensemble des règles d'association ayant une mesure de confiance supérieure ou égale à un seuil défini par l'utilisateur, appelé Min_Conf.

### *4.2. L'algorithme K-PPVBA-2T*

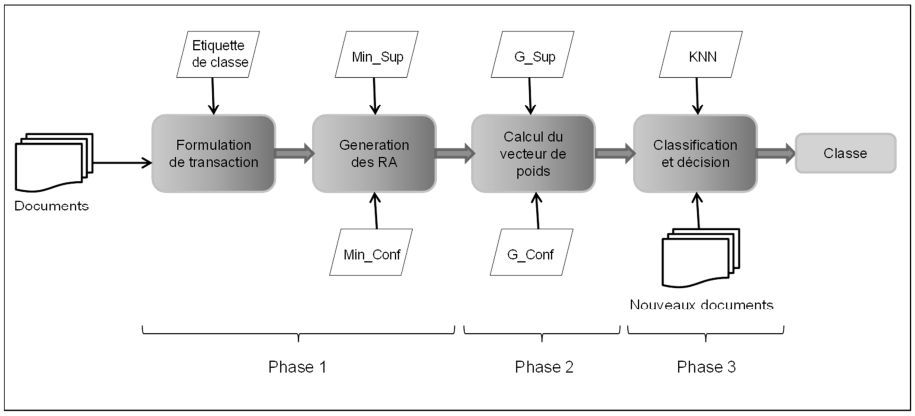


*Figure 5 : Présentation générale de l'approche K-PPVBA*

K-PPVBA est une amélioration de l'algorithme des K-PPV dont l'objectif est d'attribuer des poids à chaque attribut en utilisant les règles d'association. Nous avons utilisé les règles d'associations qui permettent d'identifier les termes les plus représentatifs d'une classe donnée. Chaque transaction est composée de l'ensemble des termes extraits (attributs) et de l'étiquette de la classe. Après la génération des règles, on ne garde que les règles de type :

$$Attribut \rightarrow Classe \text{ et } (Attribut_1, Attribut_2) \rightarrow Classe$$

Les règles composées de trois attributs sont rares et ne sont pas déterminantes.

L'idée est que si deux attributs, attribut1 et attribut2 sont associés ensembles à une classe et si chacun d'eux individuellement est associé à la même classe, ces deux attributs doivent être jugés conjointement pertinents : la force informationnelle de chacun des deux attributs déduite de leur association, est plus importante que la force informationnelle d'un attribut seul.

Nous appliquons le principe des règles d'association selon deux variations, une première (K-PPVBA-1T), où nous ne prenons en compte que les règles composées d'un seul attribut (i.e. terme), et une seconde (K-PPVBA-2T), où les règles d'un seul attribut sont déduites des règles de deux attributs.

La fonction de pondération se base sur deux paramètres : le plus grand support pour chaque attribut noté *G_Sup* et aussi la plus grande confiance pour chaque attribut appelé *G_Conf*. Par conséquent, la formule de la distance de l'algorithme K-PPV doit être modifiée en ajoutant le vecteur poids (*W*) défini comme :

$$W[i] = \left(\frac{1}{1 - G\_\mathrm{Sup}[i]}\right)$$

La nouvelle formule de calcul de distance utilisée dans la méthode K-PPVBA-2T, s'écrit alors :

$$D(a,b) = \sqrt{\sum_{i=1}^{n} W[i] \times \left(x_{ai} - x_{bj}\right)^2}$$

où *a* et *b* sont deux documents, et $x_{ai}$ et $x_{bi}$ représentent le terme *i* de chaque vecteur document.

Le processus général de l'approche K-PPVBA-2T, décrit dans la figure 5, est composé de trois phases :

Phase 1 : cette phase est constituée de deux étapes. La première étape est la construction des transactions qui représenteront les entrées pour générer les règles d'associations. Chaque document est transformé en une transaction, constituée de l'ensemble de ses descripteurs représentatifs associée à l'étiquette de sa classe. La deuxième étape est la génération des règles d'association grâce à un algorithme de recherche de type Apriori (Agrawal et Srikant, 1994).

Phase 2 : dans cette phase, nous cherchons à générer un vecteur poids pour tous les attributs de l'espace de description des documents. Pour chaque attribut, un groupe de 15 règles (15 correspondant au nombre de classes) est construit. La règle la plus pertinente (de support le plus élevé, de confiance la plus élevée) est retenue. Le vecteur poids est construit d'après la formule indiquée dans l'algorithme.

Phase 3 : cette phase consiste à appliquer l'algorithme K-PPV avec l'extension ajoutée. Pour prédire la classe d'un nouveau document par le calcul de la similarité inter-document, nous prenons en compte le vecteur poids généré dans la phase précédente.

Cette technique étend ainsi la méthode des K-plus-proches-voisins selon deux voies :

1) Tout d'abord, un schéma de pondération des descripteurs est introduit en fonction de leur poids informationnel par rapport à toutes les classes.

2) Le vote des plus proches voisins est basé sur une fonction étendue par le vecteur *w*. La seconde extension utilise la force d'activation des termes vis-à-vis de la distribution des classes.

Cette dernière extension est fondée sur l'idée que les observations de l'échantillon d'apprentissage, qui sont particulièrement proches de la nouvelle observation (*y, x*), doivent avoir un poids plus élevé dans la décision que les voisins qui sont plus éloignés du couple (*y, x*). Ce n'est pas le cas avec la méthode K-PPV : en effet seuls les k plus proches voisins influencent la prédiction, mais l'influence est identique pour chacun des voisins, indépendamment de leur degré de similarité avec (*y, x*). Pour atteindre ce but, les distances, sur lesquelles la recherche des voisins est fondée dans une première étape, sont transformées en fonction de la force (i.e. du pouvoir) du terme à activer la classe.

*Tableau 3 : Comparaison de résultats de la classification avec K-PPV et K-PPV-BA*

| | *K-PPV* | *K-PPV-BA-1T* | *K-PPV-BA-2T* |
|---|---|---|---|
| *Précision* | *0,61* | *0,65* | *0,67* |

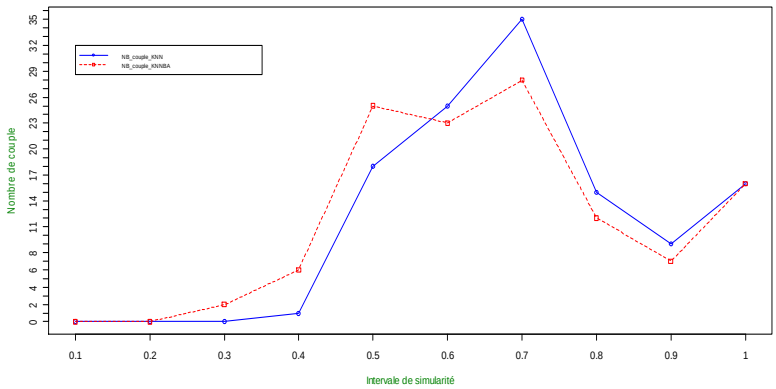


*Figure 6 : Correction du déséquilibre et de la similarité des classes avec l'approche K-PPVBA-2T*

Comme le montre la figure 6, nous avons, grâce à cette approche, joué à la fois sur la correction de la distribution des termes dans les classes et sur la correction de la similarité entre les classes. Comme le montre également la même figure, le lissage de la distribution des termes n'est cependant pas effectif sur la plus grosse classe (A61K31) qui reste toujours une classe majoritaire. Ce dernier problème handicap fortement l'évolution possible des performances.

## 5. Une nouvelle approche en perspective

L'utilisation de méthode K-PPVBA-xT présente des limitations incontournables liées au calcul des règles d'association, dont la complexité s'avère combinatoire, notamment si l'on augmente le nombre de prémisses à prendre en compte, de manière à améliorer la précision de la méthode. Dans le cadre de nos expérimentations, nous nous sommes ainsi limités à des règles à deux prémisses. De plus, le réglage des paramètres de la méthode est délicat et ses capacités de correction restent limitées, comme le montrent également nos expérimentations. Pour améliorer nos résultats, nous projetons donc de nous orienter vers l'exploitation d'autres techniques qui ne présentent pas ces types de défauts. Une technique candidate intéressante est celle des filtres détecteurs de nouveauté. Cette technique est une technique sans paramètres qui repose sur l'apprentissage incrémental à une seule classe. Comme l'ont montré (Raskutti et Kowalczyk, 2004), le principe de l'apprentissage à une seule classe permet de contourner efficacement le problème de déséquilibre des classes en considérant l'apprentissage d'une classe minoritaire comme un cas spécifique de détection de nouveauté. Le principe des filtres détecteurs de nouveauté proprement dit (NDF) a été initialement établi par (Kohonen, 1983). Il consiste à séparer l'espace de description des données en deux sous-espaces orthogonaux en apprenant incrémentalement les caractéristiques des données positives à partir d'une technique de génération séquentielle d'un projecteur d'habituation basée sur les matrices pseudo-inverses de Moore-Penrose (Penrose, 1955). Kohonen a initialement appliqué son modèle à la détection de tumeurs à partir de radios de patients. Le modèle original NDF présente cependant l'inconvénient de ne s'appliquer efficacement qu'aux cas de classes aux propriétés fortement discriminantes (Kassab et Lamirel, 2005). Pour cette raison, (Kassab et Lamirel, 2006) en ont proposé une adaptation à l'apprentissage incrémental de classes multiples aux propriétés partiellement recouvrantes, nommée ILoNDF. Cette adaptation revient à implanter une fonction d'oubli dans l'apprentissage NDF qui s'applique aux propriétés non récurrentes des classes. En s'appuyant sur des bases-test de type Reuters, il a été montré que le modèle ILoNDF était nettement plus efficace que les techniques de type SVM pour la catégorisation des données textuelles (Kassab et al., 2009). Récemment, (Hamdi et Bennani, 2011) ont proposé une méta-stratégie basée sur ce modèle, qu'ils nomment RS-NDF. Celle-ci repose sur un double bootstrap qui les amène à gérer un comité de filtres ILoNDF couvrant des sous-espaces disjoints et des données diversifiées par tirage avec remise. Pour chaque classe, les résultats du comité sont ensuite fusionnés par vote majoritaire. Comme le montrent leurs expériences, cette approche a pour avantage de relever le niveau de réponse des filtres ILoNDF en les rendant ainsi plus sélectifs, et donc plus sensibles aux caractéristiques propres aux classes, et notamment aux classes minoritaires.

Dans le cadre de notre nouvelle approche, nous projetons ainsi d'utiliser des comités de filtres ILoNDF (RS-NDF) en combinant, pour chaque classe et chaque sous-espace, des filtres d'habituation, associées aux exemples positifs, et des filtres de rejets, associés aux exemples négatifs. Dans ce contexte, ces deux types de filtres peuvent en effet avoir des périmètres d'utilisation complémentaires. Les filtres

d'habituation peuvent être utilisés pour ajuster le profil des classes en coordination avec les informations « non contradictoires[5] » de nouveauté fournies par les filtres de rejets. De manière connexe, les filtres de rejets peuvent être utilisés en coordination avec les informations de nouveauté « non contradictoires[5] » fournies par les filtres d'habituation pour réorienter l'échantillonnage éventuel, ceci de manière plus spécifique à la zone cible des données positives. Ce principe d'exploitation complémentaires des deux types de filtres est schématisé à la figure 7.

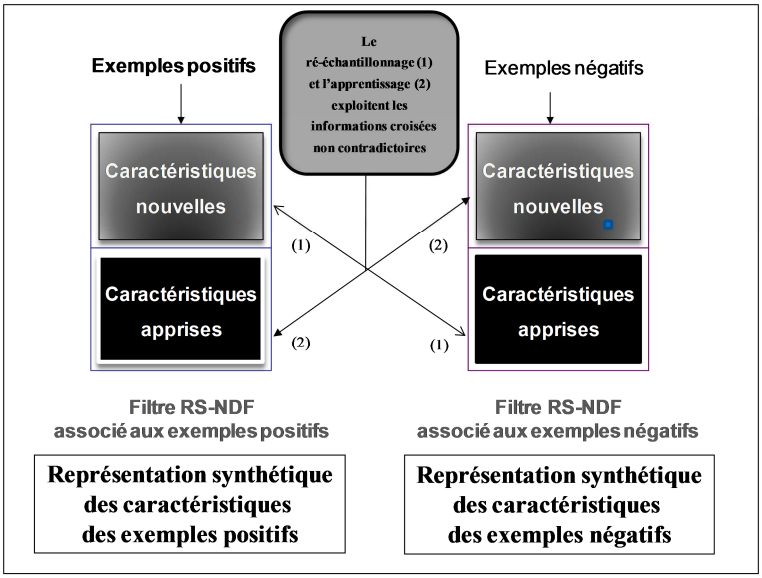


*Figure 7 : Fonctionnement des filtres détecteurs de nouveauté complémentaires d'habituation et de rejets pour l'apprentissage et le ré-échantillonnage*

## 6. Discussion et conclusion

La classification d'articles scientifiques dans un plan de classement de brevets est un véritable challenge, ce type de plan étant très détaillé et finalement peu adapté au contenu des documents scientifiques.

Dans cet article nous avons présenté une nouvelle méthode de classification supervisée issue de la méthode des K-PPV. Cette méthode que nous avons nommée K-PPV-BA-xT exploite une pondération des termes descripteurs des classes basée sur les règles d'association induites par ces termes. Nous l'avons appliqué sur un

[5] Les informations sont dites contradictoires si elles sont apprises à la fois par les filtres d'habituation et ceux de rejets, ce qui signifient qu'elles sont à la fois fortement présentes dans les exemples positifs et dans les exemples négatifs.

corpus de notices bibliographiques issues de la base Medline, dans le but de les classer dans un plan de classement de brevets du domaine de la pharmacologie. Cette nouvelle méthode offre des performances intéressantes dans notre cas d'étude. Cependant le déséquilibre et la similarité de la description des classes obtenues restent toujours des problèmes majeurs qui freinent l'amélioration des performances de la classification automatique des notices dans le plan international des brevets.

Dans le cas de la méthode que nous avons proposée, ces problèmes se cumulent avec la complexité de calcul et la difficulté de gestion des paramètres inhérents à cette dernière. C'est pourquoi, nous avons récemment entrepris de nouvelles expérimentations dans le but d'exploiter des techniques alternatives, et notamment des techniques incrémentales basées sur la détection de nouveauté, insensibles aux paramètres et possédant un meilleur potentiel pour corriger le déséquilibre des classes. Les perspectives de ce travail restent donc ouvertes.

*Remerciements: ce travail a été réalisé dans le cadre du programme QUAERO[6] financé par OSEO[7], agence nationale de valorisation de la recherche. Nous remercions Thiphaine Jadot pour l'expertise précieuse qu'elle nous a fournie concernant l'exploitation de l'outil TreeTagger dans nos expériences.*

## Bibliographie

AGRAWAL, R. et SRIKANT, R. (1994). Fast algorithms for mining association rules in large data bases. *Journal of Computer Science and Technology, Vol. 15, Issue: 6, pp 487-499.* Publisher: Morgan Kaufmann Publishers Inc.

APTE, C., DAMERAU, F. et WEISS, S.M. (1998). Text mining with decision rules and decision trees. *Proceedings of the Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web*.

COHEN, A.M. et HERSH, W.R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6:57-71*.

CORMACK, G.V. et LYNAM, T.R. (2007). Online supervised spam filter evaluation. *ACM Transactions on Information Systems, 25(3).*

DURANT, K. et SMITH, M. (2007). Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection. *In Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web (WebKDD 2006), pp 187–206.*

EVANS, M., MCINTOSH, W., LIN, J. et CATES, C. (2007). Recounting the courts? Applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies, 4(4):1007–1039.*

FONTAINE, J.F., BARBOSA-SILVA, A., SCHEFER, M., HUSKA, M.R., MURO, E.M. et ANDRADE-NAVARRO, M.A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res 37(Web Server issue), pp 141-146*.

[6] http://www.quaero.org
[7] http://www.oseo.fr/

GROVE, A., LITTLESTONE, N. ET SCHUURMANS, D. (2001), General convergence results for linear discriminant updates. *Machine Learning, 43(3): 173-210*.

HAMDI, F. et BENNANI, Y. (2011). Learning Random Subspace Novelty Detection Filters, *In Proc. IJCNN'11, IEEE International Joint Conference on Neural Network*, San Jose, California, USA.

HILLARD, D., PURPURA, S., et WILKERSON, J. (2007). An active learning framework for classifying political text. *In Annual Meeting of the Midwest Political Science Association*.

IYER, R., LEWIS, D., SCHAPIRE, R., SINGER, Y. et SINGHAL, A. (2000). Boosting for document routing. *In Proceedings of the Ninth International Conference on Information and Knowledge Management*.

JOACHIMS, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. *In Proceedings of ICML-97: 14th International Conference on Machine Learning.*

JOACHIMS, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In proceedings of the European conference on Machine learning, pp 137-142.*

KRALLINGER, M., VAZQUEZ, M, LEITNER, F, SALGADO, D et VALENCIA, A. (2010). Results of the BioCreative III (Interaction) Article Classification Task. *In Proceedings of the Third BioCreative Workshop,* Bethesda, USA.

KASSAB, R., LAMIREL, J.-C. et NAUER, E. (2005). Novelty Detection for Modeling Users Profile. *The 18th International FLAIRS Conference*, ClearWater, FL, USA.

KASSAB, R. et LAMIREL, J.-C. (2006). A new approach to intelligent text filtering based on novelty detection, *Proceedings of the 17th Australasian Database Conference (ADC 2006*), Hobart, Tasmania, AU.

KASSAB, R. et ALEXANDRE, F. (2009). Incremental Data-driven Learning of a Novelty Detection Model for One-Class Classification Problem with Application to High-Dimensional Noisy Data. *Machine Learning, 74(2): 191-234*.

KOHONEN, T. (1993). Self-Organization and Associative Memory. *3rd edition*. Springer, Berlin.

KOSTER, C.H.A., SEUTTER, M. et BENEY J. (2001). Classifying Patent Applications with Winnow, *In Proceedings Benelearn 2001*, Antwerpen, Belgia.

KOSTER, C.H.A., SEUTTER, M. et BENEY, J. (2010). Multi-classification of Patent Applications with Winnow. *In Ershov Memorial Conference 2003, pp 546-555*.

LAN, M., TAN, C.L., SU, J. et LOW, H.B. (2007). Text representations for text categorization: a case study in biomedical domain. *In: IJCNN 2007: International Joint Conference on Neural Networks.*

LEWIS, D. D. et RINGUETTE, M. (1994). Comparison of two learning algorithms for text categorization. *In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94), pp 81-93*.

LEWIS, D. D. (1992). An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. *ACM 15th Ann Int'l SIGIR'92, pp 37-50*.

MORDIAN, M. et BAARANI, A. (2009). KNNBA: k-Nearest Neighbors Based Association Algorithm. *University of Isfahan*, Iran.

PENROSE, R. (1955). A generalized inverse for matrices. *In Proceedings of the Cambridge Philosophical Society, vol. 51, pp 406-413*.

PANG, B. et LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2(1-2):1–135*.

PURPURA, S. et HILLARD, D. (2006). Automated classification of congressional legislation. *Proceedings of the international conference on Digital government research, pp 219–225*.

QUINLAN, J.R. (1986), *Induction of decision trees, Machine Learning, 1(1), pp 81-106, 1986.*

RASKUTTI, B. et KOWALCZYK, A. (2004). Extreme re-balancing for SVMs: a case study. *SIGKDD Exploration Newsletter*.

ROBERTSON, S. E., et SPARCK JONES, K. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science, 27:129–146.*

SALTON, G. et BUCKLEY, C. (1998). Term-weighting approaches in automatic text retrieval, *Information Processing Management, pp 513-523*.

SALTON, G. (1971). Automatic processing of foreign language documents. *Prentice-Hall,* Englewood Cliffs, NJ.

SCHAPIRE, R., SINGER, Y. et SINGHAL, A. (1998). Boosting and Rocchio applied to text filtering. *In Proceedings of the 21st Annual International Conference on Research and Development in Information Retrieval (SIGIR 98)*.

SCHÜTZE, H., HULL, D. A et PEDERSEN, J. O. (1995). A Comparison of Classifiers and Document Representations for the Routing Problem. *In Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR 95), pp 229-337*.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing, pp 44-49.*

SEBASTIANI, F. (1999). A tutorial on automated text categorisation, *In Analia Amandi and Ricardo Zunino, editors, Proceedings of the 1st Argentinian Symposium on Artificial Intelligence (ASAI'99), pp 7-35*.

SUOMELA, B.P. et ANDRADE, M.A. (2005). Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics, 6(75)*.

VINCARELLI, A., (2006). Indexation de documents manuscrits, *In Actes du Colloque International Francophone sur l'Ecrit et le Document (CIFED06), pp 49-53*.

WIENER, E., PEDERSEN, J.O. et WEIGEND, A.S. (1995). A Neural Network Approach to Topic Spotting. In *Symposium on document analysis and information retrieval, pp 317-332*.

YANG, Y. et CHUTE, C.G. (1994). An example based mapping method for text categorization and retrieval. *ACM Trans. Inform. Syst., 12:252-277*.

YANG Y. et LIU X., (1999). A reexamination of text categorization methods, *In Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR 99), pp 42-49.*

YIN, L., XU, G., TOTII, M., NIU, Z., MAISOG, J.M., WU, C., HU, Z. et LIU, H. (2010). Document classification for mining host pathogen protein-protein interactions. *Artif. Intell. Med 49*(3*):155-160*.

Article reçu le :AR_1religne_soumission

Article accepté le :AR_soumission

20DN.Volume 16 – n° 1/2013