PFC : Un outil d'aide à la découverte du contenu des documents et à la création de dossiers

André Alusse, Jean-Charles Lamirel, Abdel Belaïd

PFC[1] : Un outil d'aide à la découverte du contenu des documents et à la création de dossiers

André Alusse[1], Jean-Charles Lamirel[2], Abdel Belaïd[1]

1LORIA-Université Nancy 2, Campus Scientifique, B.P. 236, Vandoeuvre-Lès-Nancy, France

2LORIA, Campus Scientifique, B.P. 236, Vandoeuvre-Lès-Nancy, France

{alusse,abelaid,lamirel}@loria.fr

Résumé :

Cet article traite de la construction automatique et dynamique de dossiers consolidés. La construction de dossiers utilise plusieurs étapes : recherche des documents les plus significatifs à partir d'une requête par mots-clés, classification dynamique du résultat de la requête en utilisant plusieurs classifieurs aux comportements différentiés, combinaison des résultats de ces classifieurs pour mieux faire ressortir les thématiques extraites, et enfin personnalisation de l'organisation en introduisant les choix de l'utilisateur. Une évaluation statistique des paramètres utilisés par les classifieurs a permis de mesurer leur intérêt et surtout leurs incidences sur la constitution finale des classes thématiques. En sus de l'utilisateur, d'autres opérateurs de type plus large : groupes ou communautés peuvent intéragir avec le système pour l'enrichir. Le prototype présenté dans cet article est une plate-forme expérimentale d'observation sur l'organisation de documents et sur les méthodes de classification. L'application pilote concerne la consolidation des textes de loi de la Commission Européenne.

Mots-clés : Recherche d'information, classification automatique, combinaison de classification, dossier, partage d'informations, appropriation des documents par l'utilisateur.

1 Introduction

Cet article traite de la construction automatique et dynamique de dossiers consolidés. Nous entendons par dossier consolidé, un ensemble bien organisé de documents, et de parties de documents, traitant d'un sujet précis répondant à un besoin de l'utilisateur. Un tel type de synthèse et de consolidation d'information s'avère souvent indispensable. En effet, il intervient aussi bien dans la création d'un

[1] PFC : Personal File Consolidation

nouveau cours que dans la constitution d'un état de l'art sur une thématique, ou encore, dans des cas plus précis, comme la réalisation d'un dossier de législation à partir de différents textes de lois.

La motivation de ce projet émane principalement de demandes de SSII collaborant avec la CEE et cherchant à offrir, notamment aux juristes, la possibilité d'établir des synthèses ou des dossiers de consolidation à partir de l'ensemble des publications de la Communauté Européenne. Ces dossiers sont susceptibles de couvrir des sujets variés comme : quelles sont les dernières réglementations en matière de transport d'animaux ? Quels textes concernent plus particulièrement les douaniers en matière d'importation d'animaux et quelles sont les dernières règles en vigueur ? Quels textes concernent plus précisément les vétérinaires sur le sujet et quelle était la législation en vigueur l'année précédente ?

Au vu de ces motivations, la consolidation se traduit dans ce projet par l'intégration de plusieurs aspects du domaine de la recherche d'information : la recherche en elle même, la classification pour le rapprochement et l'organisation des documents et l'intégration du point de vue de l'utilisateur final.

A l'heure actuelle, il n'existe pas de système intégrant l'ensemble de ces fonctionnalités. De fait, les outils existants sont plutôt spécialisés sur une tâche précise : par ex. sur la recherche d'information, ou sur l'extraction d'informations et de résumés, ou sur la classification, statique ou dynamique ou bien encore, sur la catégorisation.

La catégorisation consiste à organiser les documents en se basant sur une hiérarchie thématique pré-existante. La constitution de cette hiérarchie, aussi bien que l'analyse des documents à catégoriser, sont en général des opérations manuelles laissées à la responsabilité de spécialistes.

L'objectif de la classification non supervisée est de découvrir les groupes (clusters) de documents similaires et de faire émerger des classes "latentes" d'un ensemble de documents. De nombreux types de méthodes de classification non supervisées sont proposés dans la littérature, parmi lesquels on peut citer :

- les méthodes basées sur un calcul préalable d'une matrice de similarité entre les documents, comme HAC (Hierarchical Agglomerative Clustering) [Voorhees86], Suffix Tree Clustering [Zamir98], Semantic Online Hierarchical Clustering [Zhang01] ;
- les méthodes de nuées dynamiques de type : K-means [Rocchio66] ;
- les méthodes neuronales comme les cartes auto-organisatrices de Kohonen (SOM) [Roussinov98] ou les gaz neuronaux simple (NG) [Martinetz91] ou évolutifs (GNG) [Friedske94].

Cependant, ces approches ne permettent pas d'intégrer l'expérience de l'utilisateur, ce qui est insuffisant pour la constitution de dossiers personnalisés. Il faut donc une

stratégie plus centrée autour de l'utilisateur qui prenne en compte davantage ses habitudes de travail et sa vision sur les documents. L'utilisateur doit ainsi pouvoir créer sa propre catégorisation et pouvoir la confronter à la fois à celles des autres utilisateurs et aux classifications calculées par le système.

Pour mesurer l'intérêt porté aux documents par les utilisateurs, la littérature propose la notion de "filtrage collaboratif" dont le but est de faire ressortir les avis majoritaires sur des documents qui seront ensuite utilisés comme des recommandations personnelles. Cette approche n'a cependant pas été exploitée dans le contexte plus large de l'organisation de l'information.

Parallèlement, beaucoup de travaux ont également été effectués sur la combinaison de classifications hétérogènes dans le domaine de la recherche d'information, et ceci avec des objectifs très variés. Ces combinaisons offrent des possibilités intéressantes pour fournir des résultats classifiés précis, intuitifs et complets. [Lam01] et [Bennett02] ont suggéré et généralisé de telles méthodes. Leur principale limite est cependant de fonctionner hors-ligne.

D'un autre côté, certains moteurs de recherche, comme Exalead ou Vivissimo, mettent en œuvre une classification en ligne. Mais, ils n'offrent que des possibilités réduites d'organisation qui ne répondent pas à une problématique de dossiers.

L'objectif du projet Paploo est de fournir à l'utilisateur une série d'outils pour rechercher et trier les documents en fonction de ses besoins (dossier, consolidation, synthèse, cours, etc.). Ce qui revient donc à fédérer l'ensemble des approches précédemment décrites.

L'article est organisé de la façon suivante : la solution retenue est présentée dans la deuxième section, la troisième section décrit le fonctionnement du système, la quatrième section est consacrée à la présentation des premiers résultats et à la discussion, alors que la conclusion et les perspectives sont données dans la dernière section.

2 Le système développé

2.1 La démarche

La solution de constitution de dossier que nous proposons repose sur les trois principes directeurs suivants :
1) opérer plusieurs classifications thématiques des documents, autorisant ainsi différentes visions organisationnelles des résultats,
2) exploiter l'intérêt porté par les utilisateurs sur les documents pour réaliser un filtrage beaucoup plus pertinent, ciblé sur les appréciations et besoins de ceux-ci, et de faire bénéficier l'expérience de la communauté à l'ensemble,

3) offrir différents outils d'analyse permettant à l'utilisateur de comprendre les relations entre documents et de faire ressortir les propriétés communes en conformité avec ses besoins.

2.2 L'architecture

La figure 2.1 illustre le fonctionnement de la plate-forme expérimentale PFC que nous proposons. Celui-ci se décompose en quatre étapes : la recherche (requête), les classifications, la visualisation, qui constitue la partie interactive, et enfin, la construction du dossier.

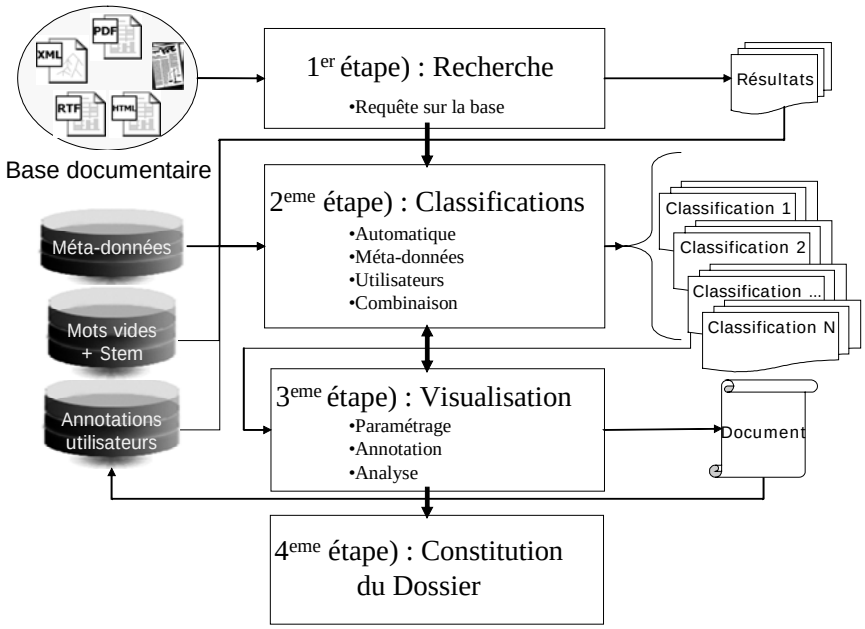


Figure 2-1 : Enchaînement des tâches dans PFC

D'un point de vue pratique, la plate-forme PFC est composée de plusieurs modules :
1) Pour la recherche : un moteur de recherche générique permettant à partir d'une requête sur un sujet de restituer les documents pertinents.
2) Pour la classification : un ensemble d'outils de classification. Ces outils sont choisis de manière à réaliser différents types de classification : agglomérative, hiérarchique, syntaxique, sémantique, ou par méta-données, afin d'obtenir plusieurs organisations de documents et un module de combinaison et d'intégration de l'ensemble de ces classifications. Chacun de ces classifieurs est interactif, permettant à l'utilisateur d'adapter ses paramètres de calcul.
3) Pour la partie interactive :

- un module d'annotation permettant à l'utilisateur de donner son propre point de vue sur les documents,
- un module d'analyse du comportement des classifieurs,
- un module de visualisation des documents.

4) Pour la constitution de dossier : un module qui réorganise les documents suivant un schéma hiérarchique choisi par l'utilisateur.

3 Description des tâches

3.1 La recherche

La requête est composée d'une suite de mots clés. En réponse à cette requête, le moteur de recherche retourne une liste de documents (D). Nous retenons le titre (T) et un extrait du document (R) en lien avec la requête, de chacun d'entre eux.

3.2 Les classifications

Classification automatique

L'objectif de la classification automatique est de regrouper les documents les plus proches à partir de la proximité des mots qu'ils contiennent. La première étape du processus se focalise sur l'extraction des mots. Chaque document est analysé de manière à constituer la liste des mots qui lui est associée. Les mots vides sont exclus et seul le radical des mots retenus (algorithme de Porter) est conservé. Pour chacun des mots retenus (radicaux), deux fréquences sont calculées : la fréquence du mot dans le document (TF pour Term Frequency) et la fréquence du mot dans l'ensemble des documents (DF pour Document Frequency). Ces valeurs permettent ensuite de calculer l'"Inverse Document Frequency" (IDF), afin de pondérer l'impact des mots sur-représentés dans les documents et qui finissent ainsi par perdre de leur valeur discriminante : si un mot est présent dans tous les documents, il ne permet pas de les différentier.

Etant donné un mot m_i, on a : $IDF(m_i) = \log((1+N) / (1+DF(m_i)))$.

Le tableau 3-1 présente un extrait de la liste des racines de mots avec le nombre d'occurrences trouvées, leur valeur de TF*IDF, ainsi que la liste des mots dont elles sont déduites.

Mot (radical)	**occurrence**	**tf*idf**	**Mot entier**
communaut	22	24,97	communauté communautés
fievr	15	33,50	fièvre
animal	72	61,01	animaux animale
utilis	5	13,20	utilisation utilise utilisés

Tableau 1 : mots extraits du vocabulaires des documents

L'objectif de l'étape suivante est d'établir une liste des séquences de mots consécutifs qui se répètent dans plusieurs documents pour obtenir des expressions

(ou motifs) les décrivant. Pour cela, on utilise la technique des "Suffix Array", introduite par [Manber93]. Le tableau 3-2 donne un extrait de cette liste.

Expression	Occurrence	Tf*idf
Protection des Animaux	4	8,93
Salaires et Rémunérations des Ouvriers d Abattoir	4	7,78

Tableau 2 : expressions extraites des documents

Puis, trois types de classifications sont finalement construites :
1) La première classification s'inspire de la méthode AHC (Agglomerative Hierarchical Clustering). A partir de ces extractions, qu'il s'agisse des mots ou des motifs, le système construit une représentation vectorielle des documents (Vector Space Model ou VSM [Salton75]). A l'issue de cette étape, chaque document sera alors décrit par les mots ou motifs associés à un coefficient de pondération égal au produit des coefficients TF et IDF. Une matrice de similarité est construite en se basant sur un calcul de distance de type corrélation cosinus [Salton75] entre les vecteurs documents. Chaque vecteur est assigné à une classe, puis une hiérarchie est établie en regroupant deux à deux les classes les plus proches, mais dont la distance est néanmoins inférieure à un certain seuil (technique du dendogramme). La classification résultante est donc un arbre dont les documents représentent les feuilles et dont les nœuds représentent les clusters. Les intitulés décrivant chaque cluster sont définis à partir des mots ou expressions majoritairement partagées par l'ensemble des documents du cluster. On peut agir sur le seuil minimal permettant le regroupement des documents et des clusters en classes.
2) La seconde classification dérive de la méthode "Suffix Tree Clustering" (STC). Les clusters sont construits à partir des mots ou des expressions les plus fréquemment retrouvés dans les documents. Un même document peut donc se retrouver dans plusieurs clusters différents. La classification n'est donc pas recouvrante. Chaque expression constitue un cluster de base. Son score dépend du nombre de mots qu'elle contient, ainsi que du nombre de documents qui la contienne. Ces clusters sont ensuite regroupés hiérarchiquement suivant leur similarité calculée à partir d'une fonction de distance.
3) La troisième classification est établie d'après la technique de Lingo [Osinski03]. Il s'agit de découvrir d'abord les labels les plus représentatifs et les plus discriminants, tout en englobant un maximum de documents. La détection de ces labels s'opère à l'aide de la technique du "Latent Semantic Indexing" [Deerwester90]. Comme pour AHC, le seuil de regroupement peut être ajusté.

Pour compléter ces classifications, nous avons implémenté trois autres techniques, dont deux sont issues du domaine neuronal. Ces techniques sont plus adaptées à la découverte de relations entre les données que les précédentes si la connaissance sur la nature de ces relations est limitée.

- La première méthode est basée sur le partitionnement de type K-Means.

– La deuxième méthode utilise les cartes auto-organisatrices SOM de Kohonen, dont le principe est le suivant : les données d'entrée sont des vecteurs, issus du VSM dans notre cas, et la carte SOM représente l'espace dans lequel elles doivent se ranger. Nous utilisons la phase d'apprentissage comme phase de découverte des relations entre documents. Les neurones de la carte (dont le nombre est fixé au départ) représentent au final les clusters de documents et les relations de voisinage entre neurones traduisent la proximité entre les classes de documents.
– La troisième méthode repose sur la technique "Growing Neural Gas" (GNG) où le nombre de neurones n'est pas imposé à l'avance comme pour les cartes SOM, ce qui permet une plus grande souplesse dans la découverte des relations possibles.

Pour ces trois méthodes, nous avons considéré que les mots et expressions représentaient les propriétés d'entrée, et que leur fréquence d'apparition dans les documents représentaient les poids des vecteurs d'entrée. Afin de restreindre l'espace d'entrée et ainsi de mieux cibler les propriétés, nous avons défini un seuil minimal de fréquence pour retenir les mots simples (S_m = 10 %), et, un seuil minimal inférieur pour retenir les expressions (S_e = 5%) afin de les valoriser.
Soit $P=\{p_1,p_2,\ldots,p_j\}$ l'ensemble des propriétés, m_i un mot et E_i une expression extraite précédemment. Nous avons donc la relation suivante :

$$m_i \in P \Leftrightarrow DF(m_i)) > S_m \text{ et } E_i \in P \Leftrightarrow DF(E_i)) > S_e$$

Soit $D = \{d_1,d_2,\ldots d_n\}$ l'ensemble des documents résultats, C le corpus. Nous avons la relation suivante :

$$C = D \times P \,/\, C_{i,j} = TF(d_i,p_j)*IDF(p_j)$$

Ce corpus ainsi constitué servira d'entrée aux trois autres méthodes de classification. A l'issue de l'application de ces méthodes, les expressions apparaissant le plus fréquemment dans les clusters serviront à constituer les étiquettes de ceux-ci.

Catégorisation par méta-données
Les vues sur l'organisation des documents obtenues précédemment sont complétées par des informations provenant des indexations thématiques établies par un expert du domaine. Chaque document est décrit avec des méta-données, comme cela est par exemple défini par la norme "Dublin Core", ou bien par le thesaurus Eurovoc, dans le cadre de notre étude. Les clusters de documents sont déduits de ces informations.

A titre d'exemple, le contenu des méta-données Eurovoc suivant, associé en parallèle à un document :
<EUROVOC_DOM CODE="56"> Agriculture, Sylviculture </EUROVOC_DOM>
<EUROVOC_MTH CODE="5641">5641 pêche</EUROVOC_MTH>
produira les deux clusters imbriqués "Agriculture, Sylviculture" et "pêche".

Catégorisation utilisateur
Au cours de la consultation, l'utilisateur est invité à donner son avis sur le document, ceci en en introduisant ses propres mot-clés pour le décrire. L'utilisateur

s'approprie ainsi les documents, et, comme ses annotations sont stockées, cela permet d'enrichir la base d'informations au fur et à mesure.
Pour construire la classification découlant de ces annotations, la liste des mot-clés décrivant chacun des documents résultats est dressée. Ce ″sac de mots″ sert à établir le VSM, point de départ d'une classification type ″AHC″.
Comme chaque utilisateur est associé à un groupe partageant le même profil et les mêmes besoins, une classification peut être déduite sur le même principe avec l'ensemble des mot-clés utilisé par le groupe. L'utilisateur peut donc enrichir ses points de vue sur les documents en découvrant les remarques d'usagers proche de lui. Une troisième classification est construite à partir des annotations de l'ensemble des intervenants. Cette approche permet à l'utilisateur de partager et confronter son avis et au système d'offrir plus de point de vue pour avoir une vision plus globale des documents.

Combinaison des classifications
La figure 3-1 illustre le mécanisme de combinaison des classifications. Chaque classification prise individuellement n'apporte pas de solution idéale. Pour bénéficier des avantages de chacune des classifications et minimiser leurs défauts, une combinaison en est donc opérée. Nous avons choisi d'utiliser à nouveau l'algorithme AHC pour effectuer cette combinaison car : 1) il est le plus simple à mettre en œuvre et 2) on peut choisir la granularité (indice) de classification recherchée pour l'application.

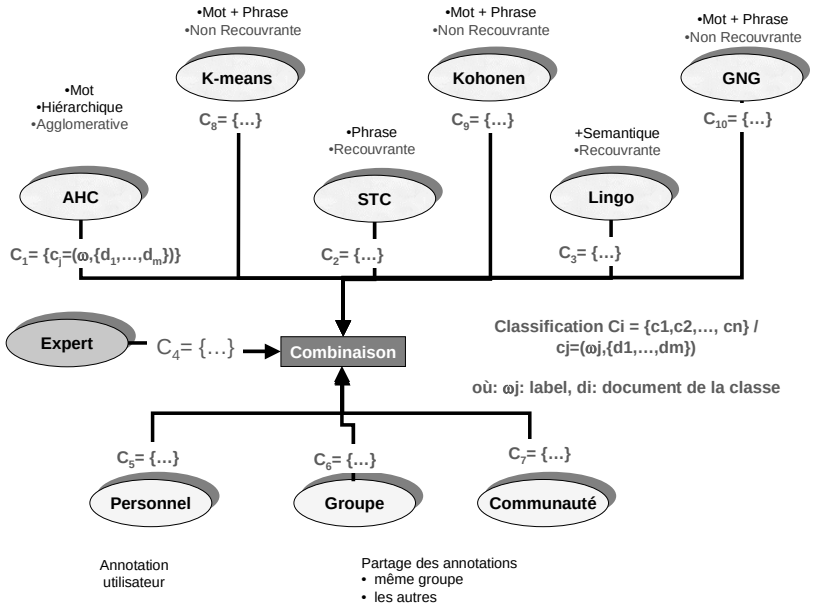


Figure 3-1 : combinaison des classifications

Soit C_i une classification, avec $C_i=\{c_1,c_2,\ldots, c_n\}$ où $c_j=(\omega_i,\{d_1,\ldots,d_m\})$; ω est le label de la classe et d_i un document de la classe. Tous les labels ω_i sont extraits des classifications. Ces labels servent ensuite à construire le VSM (M). Le poids d'un label est fonction du poids accordé au classifieur et du nombre de fois où ce label est mentionné. Le poids d'un classifieur est attribué en fonction de l'importance que l'utilisateur lui donne et de son opportunité à répondre à ses besoins.

$M_{i,j} = \sum_{k=1}^{n} (d_i \in C_k(\omega_j)) \times p_k$ où p_k est le poids de la classification k,

et $d_i \in C_k(\omega_j)$ une fonction binaire (la classe ω_j contient ou non le document d_i).

3.3 Visualisation

Il s'agit de la phase interactive du système qui permet à l'utilisateur de consulter les documents, de prendre connaissance des relations inter-documents (classification), de modifier les paramètres pour affiner le comportement des classifieurs, et enfin, d'annoter les documents.

Paramétrage des classifieurs

Comme nous l'avons préalablement mentionné, le paramétrage des classifieurs permet à l'utilisateur d'agir sur le comportement de ceux-ci et de découvrir les relations de différents niveaux de généralité entre documents. De manière pratique, ces modifications de paramètres sont prises en compte, d'abord au niveau du classifieur concerné, et ensuite, au niveau de la combinaison. Cette opération peut être itérée autant de fois que l'utilisateur le souhaite.

Annotation

En cours de consultation, l'utilisateur a la possibilité d'annoter les documents consultés, d'apporter sa propre vision sur un document, et donc d'influer sur les classifications utilisateurs, et ainsi d'enrichir le système au cours de son exploitation. De manière pratique, la note saisie (avec une interface appropriée) pour un document donné est stockée dans la base et le processus de classification est relancé pour les classifications utilisateur. Le module est en phase de réflexion et d'élaboration.

Analyse des résultats

Quatre mesures objectives ont été définies pour permettre à l'utilisateur de mesurer la qualité des classifications et de juger de l'opportunité de celles-ci pour atteindre son but. Ces mesures servent aussi à comparer le comportement des classifieurs et à découvrir les plus aptes à détecter les propriétés communes :

- la couverture permet de calculer le nombre de documents classés,
  C = (NbDocResultat – NbNonClasse) / NbDocResultat)
- la dispersion permet de mesurer la répartition en classes,
  D = NbClasse / NbDocResultat

– la précision permet de mesurer l'homogénéité d'une classe,

$$\text{Précision}\ \ P = \frac{\sum_{1}^{n} P_{C_i}}{|C|} \ / \ P_{C_i} = \frac{\sum_{1}^{n} P_{C_i}(t_j)}{|C_i|} \ / \ P(C_i(t_j)) = \frac{nbDoc(t_j \in d)}{nbDoc(Ci)}$$

– le rappel permet de mesurer l'indépendance des classes.

$$\text{Rappel}\ \ R = \frac{\sum_{1}^{n} R_{C_i}}{|C|} \ / \ R_{C_i} = \frac{\sum_{1}^{n} R_{C_i}(t_j)}{|C_i|} \ / \ R(C_i(t_j)) = \frac{nbDoc(t_j \in d, d \in C_i)}{nbDoc(t_j \in d, d \in \{C_i\})}$$

Les mesures de précision et de rappel que nous avons choisies représentent une adaptation, à l'évaluation des classifieurs, des mesures de précision et de rappel utilisées en ingénierie documentaire. Cette approche, proposée par [Lamirel04], présente l'avantage majeur d'être totalement indépendante de la méthode de classification utilisée, contrairement aux méthodes d'évaluation classiques basées sur l'inertie inter et intra classe. Elle permet donc de comparer objectivement les résultats de plusieurs classifieurs différents, ce qui s'avère important dans notre contexte.

Dossier

A l'issue de ces étapes, le dossier est créé à partir des nœuds de classification sélectionnés par l'utilisateur. Ces nœuds sont transmis à un utilitaire qui se charge de la recomposition et de la reformulation. Des considérations de recomposition et de reformulation automatique basée sur une annotation sémantique sont l'objet d'un autre travail.

4 Expérimentations et discussions

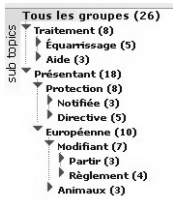
La méthode que nous avons proposée a été expérimentée avec une petite partie des documents de la Communauté Européenne, à savoir 2000 documents incluant 453 règlements, 368 questions écrites, 242 traités, etc. Ce corpus reste cependant suffisamment représentatif pour valider notre approche.

Le tableau 4.1 présente les classifications établies par les différents algorithmes pour la requête ″transport d'animaux″. Ces premiers résultats amènent les remarques suivantes :

1. pour AHC, le nombre de clusters est important ; il est difficile de juger de la qualité des clusters et de l'opportunité des termes pour l'identification des propriétés communes,
2. pour STC, les documents ne sont pas assez discriminants (beaucoup d'introductions communes à plusieurs documents) pour en déduire des regroupements intéressants. Le recouvrement est trop important,
3. Pour Lingo, moins de clusters sont détectés, les intitulés sont plus opportuns mais beaucoup de documents ne sont pas classés,
4. Pour K-means, l'équilibre entre la précision et le rappel n'est pas optimum, la taille des clusters est hétérogène,

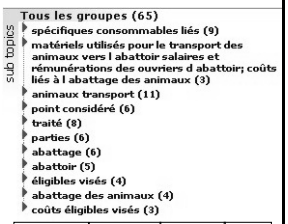
5. Pour l'approche SOM, l'équilibre est meilleur, mais au détriment d'un cluster poubelle qui regroupe 10 documents,
6. Pour l'approche GNG, les clusters sont plus équilibrés,
7. La classification issue des méta-données (non présente dans le tableau) construit beaucoup de clusters avec un fort recouvrement du fait que les documents contiennent beaucoup de méta-données.

AHC



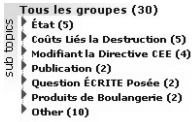
couverture	dispersion	précision	rappel
0,92	0,34	0,62	0,45

STC



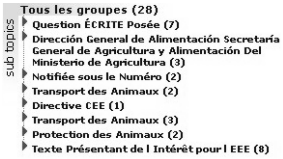
couverture	dispersion	précision	rappel
1	-	-	-

Lingo



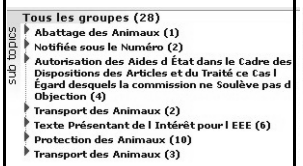
couverture	dispersion	précision	rappel
0,64	-	-	-

K-means



couverture	dispersion	précision	rappel
1	0,28	0,71	0,45

Kohonen



couverture	dispersion	précision	rappel
1	0,25	0,59	0,57

GNG

couverture	dispersion	précision	rappel
1	0,42	0,7	0,52

Tableau 3 : Tableaux des résultats de classifications

Ces différentes classifications permettent aussi à l'utilisateur de découvrir les mots et les concepts partagés par les documents. Dans le cadre des documents de la Commission Européenne, d'autres phénomènes peuvent être détectés comme par exemple : les références croisées (par ex. *vu l'article n°*…), les termes de mise à jour (par ex. *modifié le* …), les "questions écrites", les "directives", etc. Cela induit différents regroupements possibles pour l'utilisateur.

Classification utilisateurs

Afin de simuler les contributions utilisateurs, nous avons créé 8 utilisateurs virtuels, répartis en 3 groupes. Un groupe douanier avec deux profils, l'un orienté "règlement", l'autre "frontière", un second groupe vétérinaire avec trois profils, sanitaire, maladie et abattoir ; enfin un troisième groupe agriculteur ayant pour profils, agriculture, pisciculture et élevage de poulets. Pour chacun des utilisateurs, nous avons attaché des mots-clés aux documents en fonction de leurs préoccupations

supposées et d'une analyse subjective des documents, par exemple "grippe aviaire" pour le vétérinaire ayant en charge les maladies. Les résultats montrent que les clusters construits à partir de ces contributions sont en forte connections avec les préoccupations utilisateurs et donc, que l'enrichissement des données à partir des annotations utilisateurs constitue un apport primordial à notre problématique.

Analyse d'un paramètre

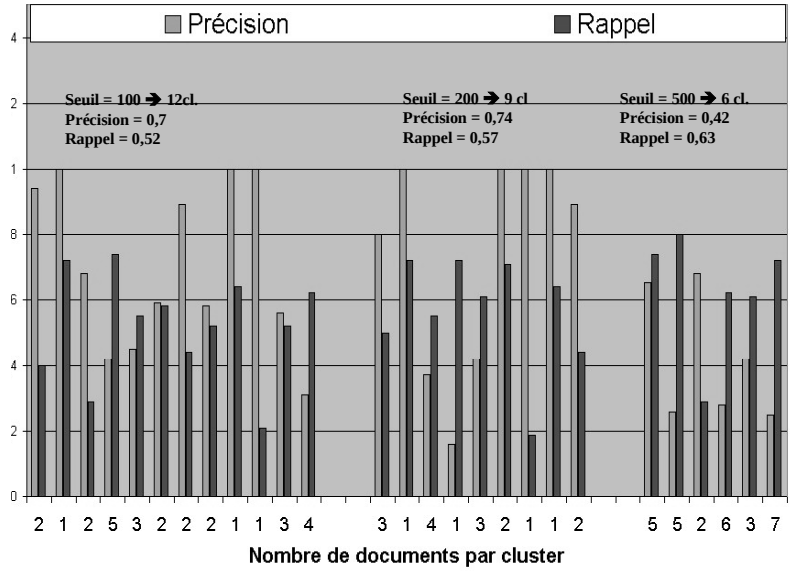


Figure 4-1 : analyse d'un paramètre pour la méthode "Growing Neural Gas"

La Figure 4-1 montre l'impact du suivi d'un paramètre dans le comportement d'une classification. L'accroissement de la valeur du seuil pour créer un nouveau nœud illustre le comportement de l'algorithme et permet d'analyser le type des clusters créés. Plus le seuil augmente, moins le nombre de clusters est élevé. Corrélativement, la taille des clusters augmente, mais la précision moyenne recule. Toutefois, les valeurs moyennes de précision et de rappel ne représentent que des valeurs indicatives. En effet, sur l'exemple de la figure 3-1, que la classification de précision moyenne la plus faible (seuil 500) a permit d'isoler un cluster de 5 documents, où il existe un équilibre, s'opérant à des valeurs élevées, entre la précision et le rappel. Ceci permet d'en déduire une forte proximité entre les documents du cluster (ici, 2 traitant de l'éradication de la peste porcine, 1 sur la fièvre aphteuse et 2 sur l'influenza aviaire). Dans la classification la plus précise en moyenne (seuil 100), ces 5 documents se trouvent répartis dans 3 clusters différents. Ce cluster potentiel de 5 documents n'a donc pas été identifié dans ce dernier cas. Cet exemple démontre bien que la possibilité pour l'utilisateur d'intervenir à tout

moment sur les seuils de classification s'avère nécessaire pour lui permettre de mieux comprendre l'organisation des documents.

Combinaison des classifications

La figure suivante analyse l'impact de chaque classifieur dans la combinaison. Pour le cas "égalité" chaque classifieur est affecté du même poids alors que dans les autres cas l'impact du classifieur analysé est multiplié par 10 afin de majorer clairement son importance. Cela nous permet de mieux comprendre la contribution de chacun des classifieurs dans la composition. La dispersion est favorisée par la contribution de l'ensemble des utilisateurs, cela étant dû à la multitude d'avis sur les documents. Elle est minimisée par les classifieurs non recouvrants, étant donné que ceux-ci discriminent moins. Le rappel est faible pour les classifieurs non hiérarchiques car ils conservent moins d'informations, à la différence de la technique AHC. Ces analyses sont primordiales pour s'approcher de la construction du dossier souhaité. Toutefois, elles sont moins évidentes à appréhender que l'impact des paramètres sur le comportement d'un classifieur donné. Elles demandent en effet une plus grande contribution de l'utilisateur, mais lui offre, en contrepartie, des outils pour une meilleure compréhension des liens possibles entre les documents, ainsi que la possibilité de contrôler l'organisation de ses dossiers.

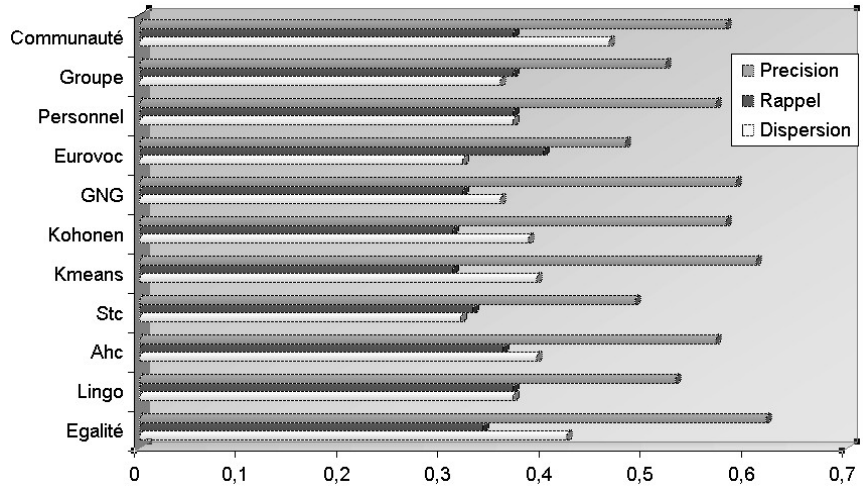


Figure 4-2 : Analyse de l'impact des classifieurs

Le tableau suivant montre trois classifications élaborées en combinant les résultats des différents algorithmes. Pour la première, le poids de chacun est le même, pour la seconde, les classifications utilisateurs et les classifications non hiérarchiques ont été privilégiées, et, dans la troisième, le poids de la classification hiérarchique a été renforcé, ainsi que les paramètres impliqués dans la hiérarchisation. Cette troisième solution permet donc d'obtenir des dossiers plus hiérarchisés alors que la seconde s'attache plus à un découpage à plat, par thème.

Ces observations permettent donc à l'utilisateur de comprendre le comportement des classifieurs et d'adapter et valider la stratégie en fonction de ses besoins.

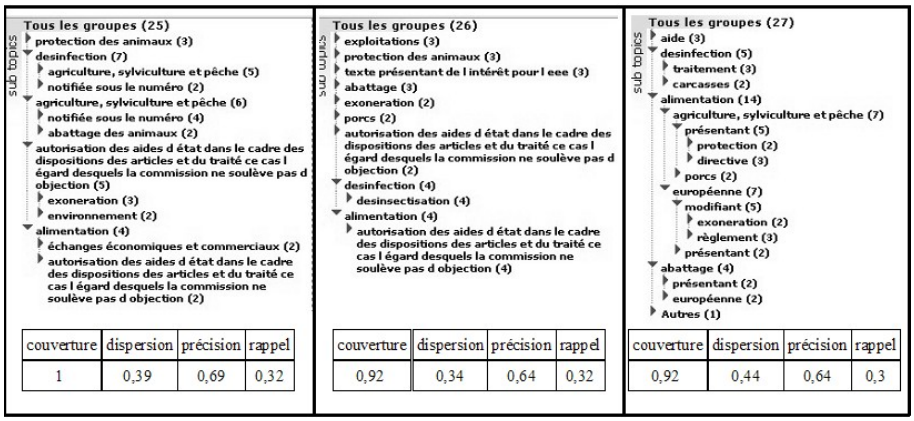


Tableau 4-2 : Exemple de résultats de combinaison de classifications

5 Conclusion et perspectives

Nous avons mis en place un nouveau système pour guider l'utilisateur dans la construction de dossier. Il en est encore au stade expérimental mais doit être progressivement intégré à la plateforme de la société leader du projet. Ce système repose sur les techniques et les idées suivantes : recherche d'information et classifications de documents, appropriation d'information et partage de connaissances, analyse d'outils et aide à la compréhension des liens et des partages de concepts entre documents. La construction de dossier passe tout d'abord par la recherche des documents les plus significatifs, puis par leur réorganisation en clusters et en hiérarchie. Ce processus interactif fait ressortir les mots, ou groupes de mots, partagés par un ensemble de documents ce qui permet à l'utilisateur de découvrir les liens existants entre documents. D'un autre côté, l'analyse des paramètres de travail et leur ajustement possible lui offre des clés pour contrôler l'organisation de son dossier. Afin de tenir compte de son profil, de sa compréhension sur les textes et de ses besoins, il est également sollicité pour enrichir la connaissance sur les documents, connaissance qui est ensuite partagée par l'ensemble de la communauté. Tout cela contribue à fournir un outil facilitant le regroupement de documents tout en tenant compte des nécessités de chaque utilisateur.

Pour améliorer le système, il est envisageable d'associer à l'utilisateur un ensemble de mots-clés ou expressions qui le caractérise, lui ou ses besoins, sous la forme d'un profil. Ces informations serviraient à renforcer l'influence des documents en adéquation avec ce profil, et par conséquent, à orienter le partitionnement vers des concepts plus en rapport avec les préoccupations de l'utilisateur. Un enrichissement de la qualification de l'impact des classifieurs et une analyse des comportements

utilisateurs dans la réalisation de l'objectif fixé pourraient permettre d'établir des stratégies et de proposer plus facilement des solutions pré-établies adaptées aux besoins particuliers. Ce travail constitue néanmoins une première approche dans l'aide à la découverte de relation entre documents et de réalisation de dossiers.

6 Bibliographie

[Bennett02] Bennett P.N., Dumais S.T. et Horvitz E., "Probabilistic combination of text classifiers using reliability indicators: Models and results", In Proceedings of SIGIR-02, Tampere, Finland, 2002.

[Deerwester90] Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R., "Indexing by Latent Semantic Analysis", J. American Society for Information Science, 1990.

[Fritzke94] Fritzke, B., Growing Cell Structures – A Self-Organising Network for Unsupervised and Supervised Learning, Neural Networks, 7(9):1441-1460, 1994.

[Lam01] Lam W. et Lai K.Y., "A meta-learning approach for text categorization", Proceedings of SIGIR-01, New Orleans, US, 2001.

[Lamirel04] Lamirel J.C., Francois C., Al Shehadi S., Hoffman M., "Multi-Topographic new classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping". In Scientometrics international Journal, Vol. 60, No. 3 445-462, 2004.

[Manber93] Manber U. et Myers G.,. "Suffix arrays: a new method for on-line string searches", SIAM Journal of Computing, 22(5), pp. 953-948, 1993.

[Martinetz91]. Martinetz T. et Schulten K., "A "neural gas" network learns topologies". In Kohonen, T., Makisara, K., Simula, O., and Kangas, J., editors, Articial Neural Networks, pages 397-402. Elsevier Amsterdam, 1991.

[Osinski03] Osinski S., "An Algoritm for Clustering of Web Search Results", Master thesis, Poznan Universitv of technology, 2003.

[Rocchio66] Rocchio J.J., "Document retrieval systems – optimization and evaluation", Ph.D. Thesis, Harvard University, 1966.

[Roussinov98] Roussinov D. et Ramsey M., "Information forage through adaptive visualization", In Proc. ACM Conf. on Digital Libraries 98 (DL98), Pittsburgh, PA, USA, 1998.

[Salton75] Salton G., Wong A., Yang C.S., "A Vector Space Model for Automatic Indexing ", Communications of the ACM, 18 (11): 613-620, 1975.

[Voorhees86] Voorhees E.M., "Implementing agglomerative hierarchical clustering algorithms for use in document retrieval", vol. 22, 465-476, Information Processing and Management, 1986.

[Zamir98] Zamir O. et Etzioni O., "Web document clustering: a feasibility demonstration", In Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), 1998.

[Zhang01] Zhang D. et Dong Y., "Semantic, Hierarchical, Online Clustering of Web Search Results", 3rd International Workshop on Web information and data management, Atlanta, Georgia, 2001.