Catégorisation automatique de textes basée sur des hiérarchies de concepts

Jérôme Augé (*), Kurt Englmeier (***), Gilles Hubert (*), Josiane Mothe (*, **)

(*) Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 04, France
(**) Institut Universitaire de Formation des Maîtres, 56 Avenue de l'URSS, 31400 Toulouse, France
(***) German Institute for Economic Research (DIW), Königin-Luise-Str. 5, 14195 Berlin,Allemagne
mothe@irit.fr, hubert@irit.fr tél: 05 61 55 63 22

RÉSUMÉ. *Cet article présente une méthode de catégorisation automatique de textes à partir de hiérarchies de concepts décrivant un domaine. Cette catégorisation se base sur deux composants essentiels :*
– la définition de représentants de catégories basée sur des principes d'apprentissage,
– un mécanisme de vote qui permet de déterminer la ou les catégories les plus adéquates pour un document donné.
Nous avons étudié l'influence de différents paramètres sur les résultats obtenus, en particulier les fonctions de choix des représentants des catégories. Les mécanismes proposés ainsi que les résultats obtenus sur la collection Reuters-21578 sont présentés dans cet article.

ABSTRACT. *This paper deals with a method for automatic categorisation of texts according to concept hierarchies that describe a domain. This categorisation is based upon two principal components:*
– the definition of category representatives resulting from learning,
– a voting mechanism in order to determine the most suitable categories for a given document.
We evaluate the influence of different parameters on the results including the methods used to select the terms to be added to the category representation. The performances that have been obtained using the Reuters-21578 corpus are reported in this paper.

MOTS-CLÉS : *Recherche d'information, hiérarchies de concepts, catégorisation automatique.*

KEYWORDS: *Information retrieval, concept hierarchies, automatic categorisation*

1. Introduction

L'organisation des informations pour permettre leur accès efficace est un problème crucial lorsque les masses d'information à gérer sont importantes. La création d'index est une des techniques d'organisation les plus utilisées par les moteurs de recherche d'information. Les contenus des documents sont analysés et des termes représentatifs (ou index) sont choisis pour chacun des documents. Lorsqu'une requête est soumise au moteur, celui-ci accède aux index correspondants et sélectionne les documents associés. Des mécanismes additionnels peuvent permettre d'ordonner ces documents par ordre décroissant de ressemblance avec la requête, par exemple lorsque les index sont pondérés. Ce type d'index est en particulier utilisé dans les moteurs de recherche sur le Web comme Google ou Altavista. La plupart des modèles de moteurs de recherche se basent sur ce type de représentation des documents selon lequel un document est considéré comme un ensemble de termes éventuellement pondérés.

Une autre méthode d'organisation consiste à regrouper les documents en fonction de leur contenu de sorte que des documents abordant les mêmes thèmes se retrouvent ensemble. Les documents qui ont des représentations suffisamment proches sont regroupés. Ce principe repose sur l'hypothèse selon laquelle les documents pertinents pour une requête ont des descriptions similaires (Rijsbergen,79). Lors d'une recherche, la requête est mise en correspondance avec une classe de documents plutôt qu'avec chaque document pris séparément. La recherche des documents pertinents pour une requête s'en trouve accélérée. Alternativement, le regroupement des documents peut aboutir à une structure hiérarchique comme dans le cas de l'utilisation d'une CAH (Classification Ascendante Hiérarchique)(Voorhees,86). Un second type de regroupement se base sur la catégorisation des documents. Dans ce cas, il s'agit d'associer à chaque document une ou plusieurs catégories prédéfinies, ces catégories pouvant être elles-mêmes organisées de façon hiérarchique. Selon ce principe, les documents associés à une même catégorie se trouvent regroupés. Du point de vue de l'utilisateur, l'accès aux documents se réalise via l'exploration de la structure des catégories qui est généralement sous forme d'un arbre. Ce principe d'organisation et d'accès aux documents est utilisé par exemple par le moteur de recherche sur le Web Yahoo! ou dans les bibliothèques numériques. On peut citer par exemple le cas des articles de MedLine auxquels les catégories du thesaurus MeSH (http://www.nlm.nih.gov/mesh/) de la National Library of Medicine sont associées ou le cas des articles de Reuters.

Il faut noter cependant que dans le cas de Yahoo! ou dans celui des collections MedLine ou Reuters, la catégorisation est réalisée de façon manuelle. Il s'agit là d'une limite importante de ce type d'organisation de l'information.

Notre approche s'intéresse à la catégorisation automatique de documents basée sur des ontologies. Il s'agit de générer automatiquement des liens entre des documents et des ontologies afin de matérialiser cette catégorisation. Dans notre

approche, les ontologies sont matérialisées par des hiérarchies de concepts (HC) qui correspondent à la connaissance d'un domaine ; les liens de catégorisation sont des liens pondérés afin de permettre la restitution ordonnée de documents lors d'une recherche. L'approche s'appuie sur l'analyse des contenus des documents et des hiérarchies ainsi que sur une méthode de vote pour définir les concepts des hiérarchies à lier aux documents. Nous avons étudié différents paramètres qui peuvent influer sur la définition des liens établis entre textes et hiérarchies prenant en compte notamment les liens sémantiques entre termes et la notion de groupes de mots. Des expérimentations nous ont permis de mesurer l'influence de ces paramètres sur la qualité des liens établis.

Cet article s'articule comme suit. Après une étude des travaux réalisés dans ce domaine, nous exposons les principes de la méthode que nous proposons pour la catégorisation automatique des textes suivant des hiérarchies de concepts. Nous présentons ensuite différentes composantes de notre approche qui peuvent influer sur la définition des liens et qui ont été pris en compte dans notre étude. Une dernière partie présente les expérimentations menées et une étude des résultats obtenus.

2. Travaux du domaine

2.1. Catégorisation automatique

La catégorisation automatique peut être réalisée efficacement dès lors qu'un apprentissage est possible. Il est alors nécessaire de disposer d'un ensemble d'apprentissage, c'est-à-dire d'un ensemble de documents pré-catégorisés. Cet ensemble sert de référence pour paramétrer l'outil de catégorisation. Lorsque l'apprentissage est terminé, de nouveaux documents peuvent alors être catégorisés, en fonction de leur adéquation avec les catégories apprises. Différentes techniques issues de la statistique ou de l'apprentissage automatique se basent sur ce principe. Récemment, les algorithmes de type SVM (Support Vector Machine) (Vapnik,95) ont été introduits pour catégoriser des documents. Ce type d'algorithme a montré son efficacité aussi bien dans le cas de catégorisation plane (Joachims,98), que dans le cas de catégorisation hiérarchique (Dumais,00). Néanmoins, dans le cas de grandes collections, la complexité de ces algorithmes devient trop importante et nécessite l'introduction de méthodes d'optimisation liées au problème de programmation quadratique (Chin,99). Dans le cas de gros volumes de documents, la simplicité du modèle de Rocchio peut s'avérer intéressante (Vinot,02). Le modèle de Rocchio a été utilisé initialement dans le cadre de la reformulation de requête par réinjection de pertinence (Rocchio,71). Il a été adapté pour permettre la catégorisation avec phase d'apprentissage. Ce modèle se base sur une représentation des documents et des catégories sous forme de vecteurs dans l'espace des termes d'indexation (Salton,71). Lors de l'apprentissage, les poids du vecteur de la

catégorie (ou profil) à laquelle appartient le document d'apprentissage sont modifiés. Les exemples positifs contribuent positivement au profil de la catégorie (augmentation des poids des termes associés à ces documents) alors que les exemples négatifs contribuent en sens inverse. De nombreuses adaptations de ce modèle ont été proposées, par exemple en interdisant des poids négatifs dans la représentation des catégories, en normalisant les vecteurs obtenus ou en ajustant l'importance relative des exemples positifs et négatifs, soit globalement, soit en fonction des catégories (Lewis,96), (Joachims,97). D'autres modèles issus de la statistique ou de l'apprentissage automatique sont utilisés pour la catégorisation ou la classification automatique comme les modèles Bayésiens (Koller,97), (Lewis,94), les modèles à régression multi-variée ou les réseaux de neurones (Schutze,95).

2.2. Utilisation de la catégorisation pour l'accès aux documents

Différents systèmes exploitent l'organisation des documents selon des hiérarchies de concepts ou de catégories pour offrir des moyens d'accès aux documents.

Cat-a-Cone (Hearst,97) est une interface graphique 3D de recherche de documents suivant des hiérarchies de catégories. Dans Cat-a-Cone les documents peuvent être associés à plusieurs catégories d'une hiérarchie, par exemple la hiérarchie Mesh. Les hiérarchies sont présentées à l'utilisateur sous forme d'arbres 3D. La construction de requêtes peut être réalisée en sélectionnant les catégories au niveau de l'arbre et en étiquetant les catégories sélectionnées par différentes couleurs pour spécifier les conjonctions ou disjonctions. Les documents associés sont représentés sous forme de livre regroupant le titre du document, les catégories associées au document et le contenu du document. Lors de la visualisation d'un document la représentation 3D de la hiérarchie indique automatiquement l'espace de concepts dans lequel se trouve le document.

Le système développé dans le cadre du projet IRAIA (IRAIA) se base sur le même principe de catégorisation multiple. Il s'agit en fait d'une catégorisation par facette où chaque hiérarchie correspond à un aspect ou une facette du domaine. Par exemple, le domaine de l'économie peut être structuré suivant les hiérarchies que sont les régions, les industries et les indices économiques. Les hiérarchies sont utilisées pour catégoriser les documents et constituent également le langage d'interrogation du système. Les hiérarchies sont présentées à l'utilisateur sous forme d'arbre. L'utilisateur interroge le système par navigation dans ces arborescences en sélectionnant les catégories correspondant à son besoin. Le système fournit alors la liste ordonnée des documents associés aux catégories sélectionnées. L'utilisateur peut ensuite visualiser le contenu des documents ainsi que l'ensemble des catégories qui leurs ont été associées. La nouvelle requête peut alors être soumise au système pour produire un nouveau résultat. Dans ce type de système, l'utilisateur reste dans un même contexte sémantique lors de sa recherche car il ne peut pas rencontrer par hasard des données qui appartiennent à un autre contexte (Englmeier,01). Le fait de grouper les données en fonction d'espaces identifiés résout le problème d'ambiguïté sur lequel les moteurs de recherche traditionnels butent.

Le système DocCube (Mothe,02) met également à profit une catégorisation multi-facette. Cependant dans ce cas, plus qu'un système de recherche d'informations, il s'agit d'un système d'exploration de grosses collections de documents. Comme dans IRAIA, l'espace d'information est structuré suivant des hiérarchies de catégories et les documents sont associés à ces hiérarchies. Un des aspects originaux de DocCube est qu'il repose sur une modélisation multidimensionnelle, qui permet de proposer à l'utilisateur des visualisations globales d'information l'aidant dans sa recherche et dans l'exploration de la masse d'information. Comme dans les systèmes OLAP –On-Line Analytical Processing–, l'information est donc représentée et organisée selon différentes dimensions et des faits peuvent être analysés de façon interactive (Chaudhuri,97). Les dimensions correspondent aux hiérarchies alors que le fait analysé est le nombre de documents associés aux nœuds des hiérarchies. Les opérateurs de forage permettent à l'utilisateur de choisir de façon interactive le niveau d'agrégation (de spécificité) auquel il veut analyser les informations. Les opérateurs de coupes quant à eux permettent d'explorer la collection par rapport à deux dimensions uniquement et d'accéder aux documents associés à un nœud.

3. Méthode de vote pour la catégorisation de textes

La catégorisation des textes repose sur le rattachement des documents à des catégories prédéfinies. Dans notre approche, nous nous intéressons à des catégories organisées selon une structure arborescente. La catégorisation de textes peut être vue comme l'association automatique des textes à différents nœuds d'une hiérarchie de catégories (HC).

3.1. Description générale de la méthode

Contrairement aux mécanismes de classification qui permettent de déterminer la classe d'appartenance d'un texte, dans notre approche, chaque texte peut être associé à plusieurs catégories d'une hiérarchie. L'association d'un texte aux catégories repose sur la méthode Vector Voting (Pauer,2000). Cette méthode se base sur la prise en compte des termes représentatifs de chaque catégorie et sur leur extraction automatique dans le contenu du texte à catégoriser. L'importance de l'association du texte avec une catégorie donnée est calculée par une méthode de vote, qui peut être rapprochée de la méthode HVV (Hyperlink Vector Voting) utilisée dans le contexte du Web pour calculer la pertinence d'une page en fonction des sites qui y réfèrent (Li,1998). Dans notre contexte, plus les termes associés à la catégorie sont présents dans le texte, plus le lien entre le texte et cette catégorie sera fort.

3.2. Etapes de la méthode

Le principe d'association d'un document à des catégories regroupe différentes étapes :

– Calcul du profil de chaque catégorie d'une hiérarchie. Cette étape se base sur un apprentissage à partir de documents déjà catégorisés.

– Extraction automatique des concepts représentatifs du document et de leur importance au sein du document. Le processus d'extraction est basé sur un ensemble de règles qui utilisent par exemple les balises des documents, complété par des fonctions syntaxiques et sémantiques pour gérer les synonymes et l'élimination de termes inintéressants.

– Pour chaque catégorie de la hiérarchie, calcul d'un score selon une méthode de vote qui mesure la représentativité de la catégorie pour le texte. Le calcul de score peut être basé sur différentes fonctions de calcul qui peuvent faire intervenir des mesures comme l'importance d'un terme dans le texte et dans la hiérarchie, la taille du texte et de la hiérarchie, le nombre de termes décrivant une catégorie présents dans le texte.

– Ordonnancement des catégories de la hiérarchie dans l'ordre des scores obtenus, puis sélection de l'ensemble des catégories à associer au texte suivant une stratégie définie (par exemple, scores supérieurs à un seuil donné, ou les *n* premiers meilleurs scores).

3.3. Fonction de calcul

Différentes fonctions de vote ont été étudiées et les résultats détaillés sont présentés dans (Augé,01). La fonction de vote doit tenir compte de l'importance de chacun des termes de la catégorie dans le document, du pouvoir discriminant de chaque terme de la catégorie, de la représentativité de la catégorie dans le document. La fonction qui a donné les meilleurs résultats s'exprime de la façon suivante :

$$Vote(E_H, D) = \sum_{\forall t \in E} \frac{F(t,D)}{S(D)} \cdot \frac{S(H)}{F(t,H)} \cdot e^{\frac{NT(E,D)}{NT(E)}} \qquad (1)$$

où E_H correspond à la catégorie E dans la hiérarchie H, D est un document

$\frac{F(t,D)}{S(D)}$ où F(t,D) correspond au nombre d'occurrences du terme t dans le document D et S(D) correspond à la taille (nombre de termes) de D. Ce facteur mesure donc l'importance du terme t dans le document D.

$\frac{S(H)}{F(t,H)}$ où F(T,H) correspond au nombre d'occurrences du terme t dans la HC H et S(H) correspond à la taille de H. Ce facteur mesure donc l'importance du terme t dans la hiérarchie H c'est-à-dire son pouvoir discriminant.

$\frac{NT(E,D)}{NT(E)}$ où NT(E) correspond au nombre de termes de la catégorie E et NT(E,D) correspond au nombre de termes de la catégorie E qui apparaissent dans D. Ce facteur mesure donc le taux de présence des termes représentatifs de la catégorie (l'importance de la catégorie) dans le texte.

La fonction (1) accorde la même importance aux deux facteurs que sont l'importance d'un terme dans le texte et l'importance de ce terme dans la HC. L'application de la fonction exponentielle au taux de présence des termes représentatifs d'une catégorie dans le texte a pour but d'accentuer l'importance de cette occurrence. Une prépondérance est accordée à la proportion des termes d'une catégorie présents dans le texte.

4. Représentation des catégories

Une hiérarchie de catégories est composée de nœuds (ou catégories) et de relations "est-un" entre eux. Une catégorie est définie par un profil. En classification automatique, les profils correspondent généralement à un ensemble de termes pondérés (Salton,71), (Rijsbergen,79). Ces profils peuvent être définis lors de la phase d'apprentissage durant laquelle la connaissance d'exemples de documents bien classés permet de décider de la représentation optimale des catégories. L'association automatique des textes aux catégories est basée sur le calcul de la ressemblance entre le profil de chaque catégorie et la représentation du texte à catégoriser.

Dans notre étude, nous avons considéré différents mécanismes pour définir les représentants de catégorie.

4.1. Représentation initiale

Une catégorie est initialement représentée par l'ensemble des termes qui composent le libellé de la catégorie dans la hiérarchie. Il s'agit là d'une représentation assez pauvre qui peut être enrichie par différentes techniques, dont l'apprentissage.

4.2. Utilisation de relations inter ou intra catégories

Une première possibilité consiste à considérer la nature hiérarchique des catégories. Dans ce cas, une catégorie peut être représentée par l'ensemble des termes du nœud correspondant, enrichi par les termes des pères de ce nœud. Le niveau de profondeur peut être variable, c'est-à-dire qu'il est possible de ne

considérer que le parent direct ou de remonter jusqu'à la racine de la hiérarchie. Les liens hiérarchiques sont des liens "est-un" entre termes.

D'autres types de relations peuvent être prises en compte au travers par exemple, de la lemmatisation ou la radicalisation qui peuvent permettre de limiter les problèmes de variantes des termes pour représenter un même concept. L'algorithme de Porter (Porter,80) est utilisé dans ce sens pour les textes en anglais. D'autres algorithmes existent dans la littérature (Frakes,92), et pour différents langages (par exemple Omstem du projet Omseek (http://omseek.sf.net) permet de radicaliser les termes issus de textes en anglais, allemand, italien et français).

La représentation de textes sous la forme d'ensembles de termes simples ou de groupes de mots reste une question ouverte en RI (TREC). Il existe différents moyens de détecter des groupes de mots dans un texte (Mothe,00). Dans notre approche, nous avons défini la notion de couverture pour tenir compte de l'appartenance des termes à des groupes de mots représentant chacune des catégories. L'objectif de la couverture est de s'assurer que les catégories représentées par des termes dont peu apparaissent dans le document ont un score nul et ne font pas partie de l'ensemble des catégories associées au document. Le seuil au-delà duquel une catégorie sera retenue est calculé par rapport à un pourcentage (**couverture**) de termes constituant la représentation de la catégorie et apparaissant dans le texte. Une couverture de 50% indique qu'au moins la moitié des termes associés à la catégorie doivent être présents dans le texte pour que celle-ci soit retenue.

4.3. Apprentissage

L'apprentissage de nouveaux termes est un moyen d'enrichir la description d'une catégorie par l'adjonction de termes sémantiquement liés aux termes initiaux. Dans ce cas, ce sont des liens de type "associé à" qui sont détectés.

Si des documents sont initialement catégorisés dans la hiérarchie, l'analyse des textes peut fournir des indicateurs sur les termes qu'il serait judicieux d'ajouter à la représentation des catégories pour améliorer la catégorisation de nouveaux textes. Cette analyse peut se baser sur des éléments statistiques comme les fréquences des termes dans les documents d'une catégorie donnée, leur fréquence dans l'ensemble d'apprentissage, etc.

Nous avons étudié différentes fonctions de choix des termes extraits pour enrichir les représentations des catégories. Ces fonctions se basent sur des fonctions de poids d'indexation de la littérature en recherche d'informations (Ponte,98), (Rijsbergen,79), (Salton,71), (SparkJones,76), (TREC7). La définition de ces fonctions est issue de travaux débutés par les Professeurs D.Harper et J.Mothe à la School of Computer and Mathematical Sciences d'Aberdeen en 2000.

La fonction ayant donné les meilleurs résultats est la suivante :

L1freqngva : $R = P_i - Qn_{gi}$ où $P_i = \frac{Est(Fi_g)}{Est(TF_g)}$ et $Qn_{gi} = \frac{Est(Fin_g)}{Est(TFn_g)}$

en utilisant les composants suivants :

nb = Nombre de termes distincts dans un groupe de documents Fi_g = Fréquence du terme dans le groupe Fi_c = Fréquence du terme dans la collection $Fin_g = Fi_c - Fi_g$ $TF_g = \sum Fi_g$ $TF_c = \sum Fi_c$	$Est(Fi_g) = Fi_g + 0.5$ $Est(Fin_g) = Fin_g + 0.5$ $TFn_g = TF_c - TF_g$ $Est(TF_g) = TF_g + \frac{nb}{2}$ $Est(TFn_g) = TFn_g + \frac{nb}{2}$

Quelle que soit la fonction utilisée, le principe reste le même : un poids est associé à chaque terme qui peut être ajouté à la description d'un nœud, puis les termes sont classés dans l'ordre décroissant de poids et enfin les X meilleurs termes sont ajoutés à la description du nœud.

Dans la suite de cet article, seuls les résultats obtenus en utilisant cette "meilleure" fonction sont présentés.

5. Evaluation

5.1. Collection Reuters-21578

La procédure de catégorisation a été évaluée sur la collection Reuters-21578 (http://www.daviddlewis.com/resources/testcollections/reuters21578/). Il s'agit d'une des collections les plus utilisées pour évaluer les techniques de catégorisation et de classification (Lewis,96), (Dumais,98), (Joachims,98). La qualité de cette collection a été améliorée par David Lewis afin de supprimer des formatages ambigus, les documents doubles, régler des erreurs typographiques et de mauvaises catégorisations. La collection ainsi modifiée s'appelle "ModLewis" et regroupe une collection d'apprentissage (13625 documents), une collection de test (6188

documents) et 135 catégories. L'apprentissage est réalisé sur la collection d'apprentissage, ses performances sont évaluées sur la collection de test.

5.2. Critères d'évaluation

Les critères d'évaluation que nous avons utilisés sont directement issus des critères utilisés pour évaluer les systèmes de recherche d'informations : les taux de rappel et de précision. Dans notre étude, ces taux sont définis par :

$$Taux\ de\ rappel = \frac{Nombre\ de\ catégories\ retrouvées\ et\ \ pertinentes}{Nombre\ de\ catégories\ \ pertinentes}$$

$$Taux\ de\ \ précision = \frac{Nombre\ de\ catégories\ retrouvées\ et\ \ pertinentes}{Nombre\ de\ catégories\ retrouvées}$$

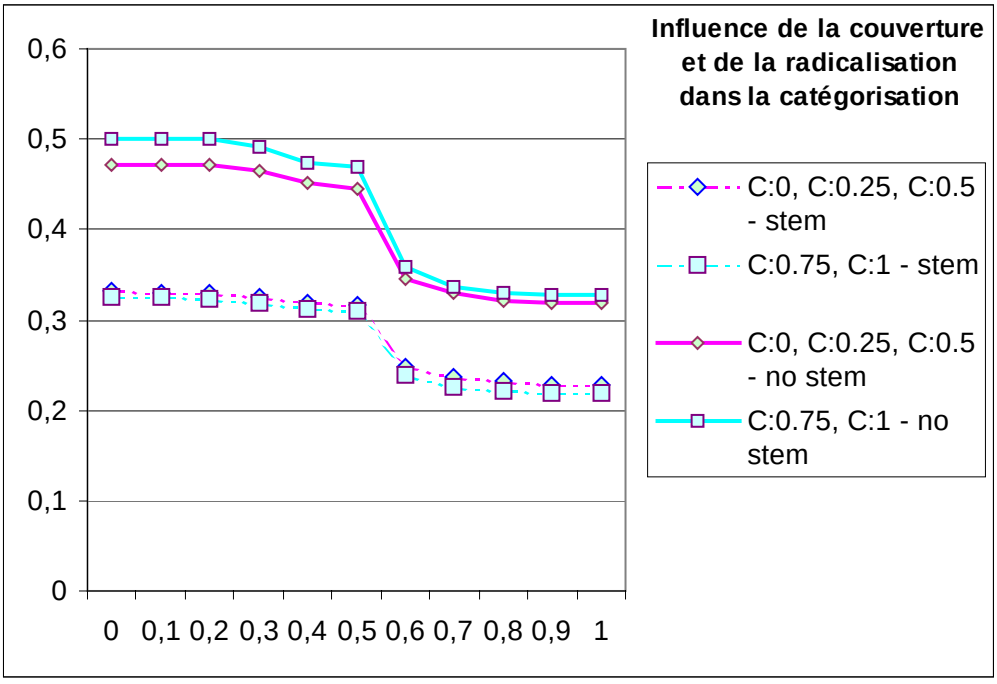
Nous avons utilisé le programme trec-eval (cf. le site web trec.nist.gov) pour calculer les valeurs de rappel et de précision. Le taux de rappel et le taux de précision évoluent généralement de manière opposée. Le taux de précision est habituellement calculé pour différentes valeurs du taux de rappel (0, 0,1, 0,2, …, 1). La R-précision est également indiquée ; elle mesure la précision après R catégories retrouvées où R est le nombre total de catégories pertinentes pour un document.

5.3. Résultats et Discussion

Les résultats obtenus pour les fonctions évaluées sont présentés au travers de figures synthétisant les mesures de rappel/précision. Chaque figure détaille la précision à différentes valeurs du taux de rappel. La R-précision est également indiquée.

5.3.1. Représentation initiale

La figure 1 regroupe les courbes qui décrivent les taux de précision à différents niveaux de taux de rappel pour différents taux de couverture (0, 0.25, 0.5, 0.75 et 1) avec et sans radicalisation. Quel que soit le taux de couverture, les résultats sont comparables voire identiques (par exemple pour les taux de couverture 0, 0.25 et 0.5). Cela peut s'expliquer par le fait que plus de la moitié des catégories ne sont initialement représentées que par un seul terme (cf. figure 2). La couverture n'a pas d'influence sur l'attachement des documents à ces catégories. Pour la même raison (représentation pauvre des catégories), les taux de précision restent faibles, quels que soient les taux de rappel, puisqu’ils sont inférieurs à 0.35.


Influence de la couverture et de la radicalisation dans la catégorisation
0,6
0,5
0,4
0,3
0,2
0,1
0
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1
C:0, C:0.25, C:0.5 - stem
C:0.75, C:1 - stem
C:0, C:0.25, C:0.5 - no stem
C:0.75, C:1 - no stem

		Couverture				
		0	0,25	0,5	0,75	1
R-Précision	Avec radicalisation	0,2388	0,2388	0,2388	0,2387	0,2387
	Sans radicalisation	0,3758	0,3758	0,3758	0,4066	0,4066

Figure 1. *Performances pour l'ensemble de test, avant apprentissage, avec et sans radicalisation des termes*

L'influence de la radicalisation (algorithme de Porter (Porter,80)) peut être observée sur la figure 1 en comparant les courbes de test avec radicalisation (stem) et sans radicalisation (no stem) pour les mêmes taux de couverture. Lorsque les termes sont radicalisés, la R-Précision est inférieure à 0,24. Le taux de précision varie de 0,21 (rappel 1) à 0,33 (rappel 0,1).

Lorsque les termes ne sont pas radicalisés, la précision varie de 0.37 à 0.40, suivant le taux de couverture. La précision à différents niveaux de taux de rappel se situe entre 0.31 (à un taux de rappel de 1) et 0.50 (pour un taux de rappel faible). Ainsi, le bruit (catégories non pertinentes associées) amené par la radicalisation n'est pas compensé par la diminution du silence (catégories pertinentes non associées).

Ces résultats correspondent aux résultats de base. Les résultats après apprentissage ne peuvent pas être comparés en relatif (les résultats initiaux étant très bas compte tenu de la pauvreté de la description initiale des catégories). En revanche, il permettent d'analyser les évolutions, en particulier par rapport aux autres paramètres (radicalisation, couverture, taille de la représentation des catégories).

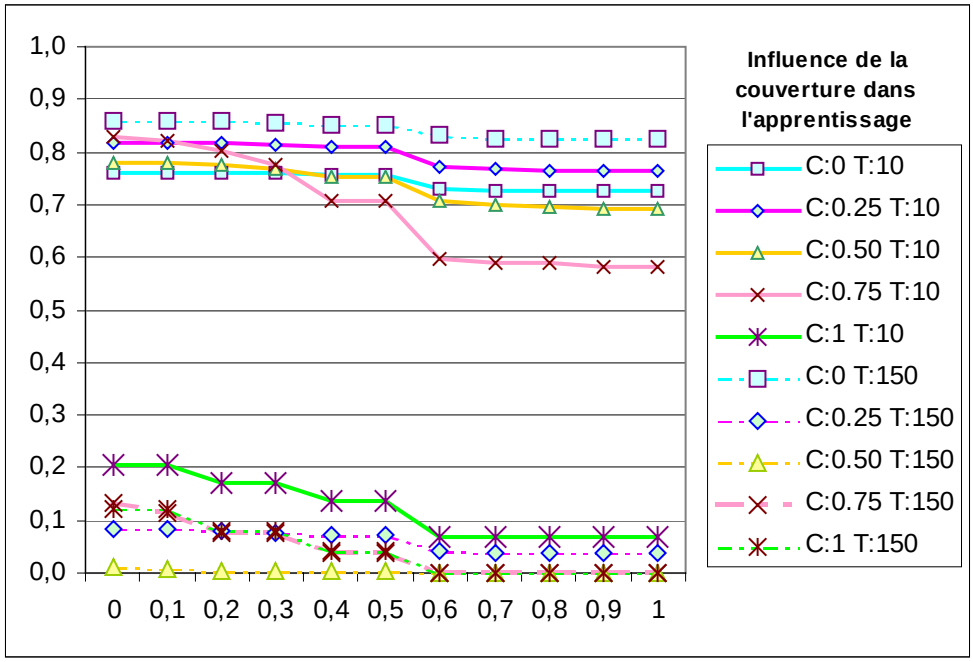
5.3.2. Apprentissage

Les meilleurs résultats relatifs à l'apprentissage ont été obtenus en utilisant la fonction L1freqngva (cf 4.3.). Seuls les résultats obtenus en utilisant cette fonction sont présentés. La comparaison avec la fonction initiale sans apprentissage est également présentée.

5.3.2.1. Influence de la couverture dans l'apprentissage

L'hypothèse établie est que plus le nombre de termes représentant une catégorie augmente plus les chances de satisfaire une couverture importante diminuent. La couverture doit donc augmenter inversement à la taille de la représentation des catégories.

La figure 2 présente les graphes de taux de précision en fonction de taux rappel ainsi que la R-Précision pour respectivement 10 et 150 termes dans la représentation des catégories. Ces résultats ont été obtenus sans radicaliser les termes afin que seule la modification du paramètre couverture influe sur les résultats.


1,0
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0,0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1
Influence de la couverture dans l'apprentissage
C:0 T:10
C:0.25 T:10
C:0.50 T:10
C:0.75 T:10
C:1 T:10
C:0 T:150
C:0.25 T:150
C:0.50 T:150
C:0.75 T:150
C:1 T:150

	Représentation des catégories	Couverture				
		0	0,25	0,5	0,75	1
R-Précision	10 termes	0,6514	0,7450	0,7205	0,6773	0,1207
	150 termes	0,7771	0,0578	0,003	0,0417	0,04

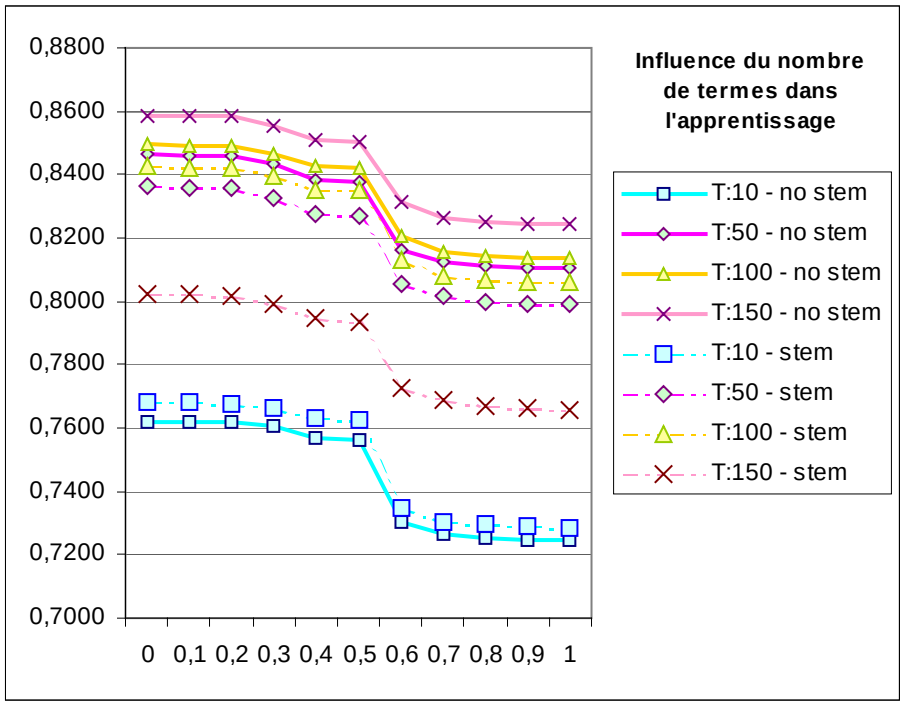
Figure 2. *Performances basées sur une représentation à 10 termes et 150 termes avec différents taux de couverture et sans radicaliser les termes*

Les résultats confirment l'hypothèse selon laquelle la couverture doit augmenter inversement à la taille de la représentation des catégories, notamment pour une représentation à 150 termes. Lorsque la taille de la représentation de chaque catégorie est importante, il est évident qu'une couverture de 100% ne peut donner de bons résultats. En revanche, il est intéressant de noter pour une représentation à 10 termes, qu'une couverture de 0,25 donne de meilleurs résultats qu'une couverture de 0.

Les meilleurs résultats ont été obtenus pour un nombre de termes à 150 avec une couverture nulle, c'est-à-dire lorsqu'aucune contrainte sur le pourcentage de termes de la représentation des catégories n'est appliquée. Dans ce cas, le taux de précision est supérieur à 0,82 quel que soit le taux de rappel. La R-précision est alors de 0,77. La R-Précision est augmentée de 4% lorsque la taille de la représentation passe de 10 termes (la R-Précision est alors de 0,74 avec une couverture de 0,25) à 150 termes (R-Précision de 0,77 pour une couverture de 0). Dans le cas d'une représentation à 150 termes, la catégorisation de nouveaux documents est bien sûr plus coûteuse en calculs, mais la précision des résultats est meilleure.

5.3.2.2. Influence du nombre de termes dans la représentation

Pour évaluer l'influence du nombre de termes dans la représentation des catégories indépendamment du paramètre couverture, nous avons fixé la valeur de la couverture à 0.


Influence du nombre de termes dans l'apprentissage
0,8800
0,8600
0,8400
0,8200
0,8000
0,7800
0,7600
0,7400
0,7200
0,7000
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1
T:10 - no stem
T:50 - no stem
T:100 - no stem
T:150 - no stem
T:10 - stem
T:50 - stem
T:100 - stem
T:150 - stem

		Nombre de termes			
		10	50	100	150
R-Précision	Sans radicalisation	0,6514	0,7637	0,7654	0,7771
	Avec radicalisation	0,6625	0,7451	0,7508	0,7022

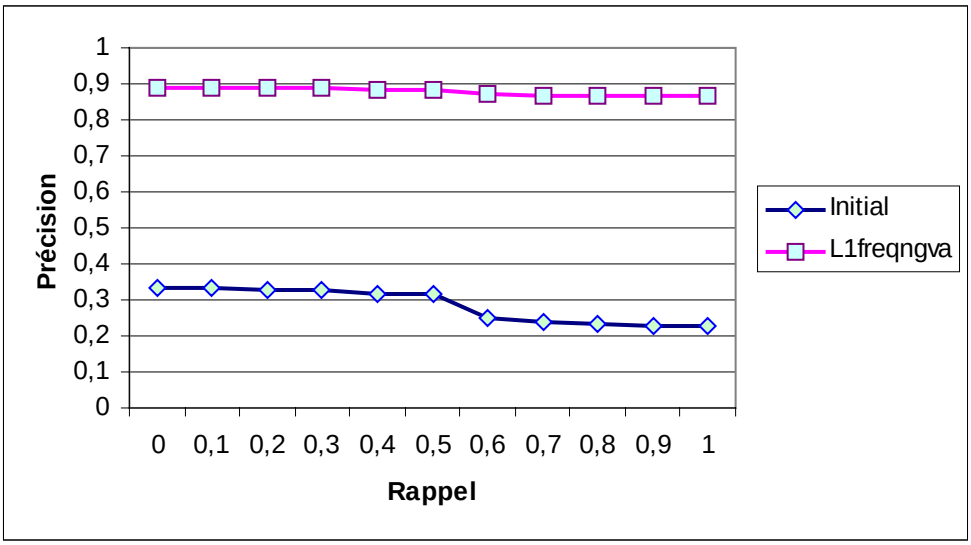
Figure 3. *Performances en fonction de la taille des représentations des catégories, avec et sans radicalisation*

Lorsque les termes sont radicalisés, la taille maximale de la représentation des catégories se trouve entre 100 et 150. Cette taille peut être supérieure lorsque les termes ne sont pas radicalisés. Le fait que la taille maximale soit atteinte d'abord avec des termes radicalisés correspond à un résultat intuitif. Les termes rajoutés après 100 termes sont marginaux pour représenter la catégorie.

Globalement, les meilleurs résultats sont obtenus sans radicalisation. Par exemple, avec la radicalisation la meilleure R-Précision est de 0,75 (pour une représentation à 100 termes) alors qu'elle est de 0,77 sans radicalisation (et pour une représentation de 150 termes). Il est intéressant de noter que les variations en terme de performance sont faibles lorsque l'on augmente la taille des représentations au-delà de 50 termes. Dans le cas où la radicalisation n'est pas appliquée, la R-Précision est augmentée de 0,2% entre 50 et 100 termes et de 1,7% entre 50 et 150.

5.3.2.3 Collection de test

Les meilleurs résultats sur la collection d'apprentissage ont été obtenus en utilisant la fonction L1freqngva avec une couverture de 0 et 100 termes ajoutés. Les paramètres correspondant à ces résultats ont donc été appliqués à la collection de test.


Précision
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1
Rappel
Initial
L1freqngva

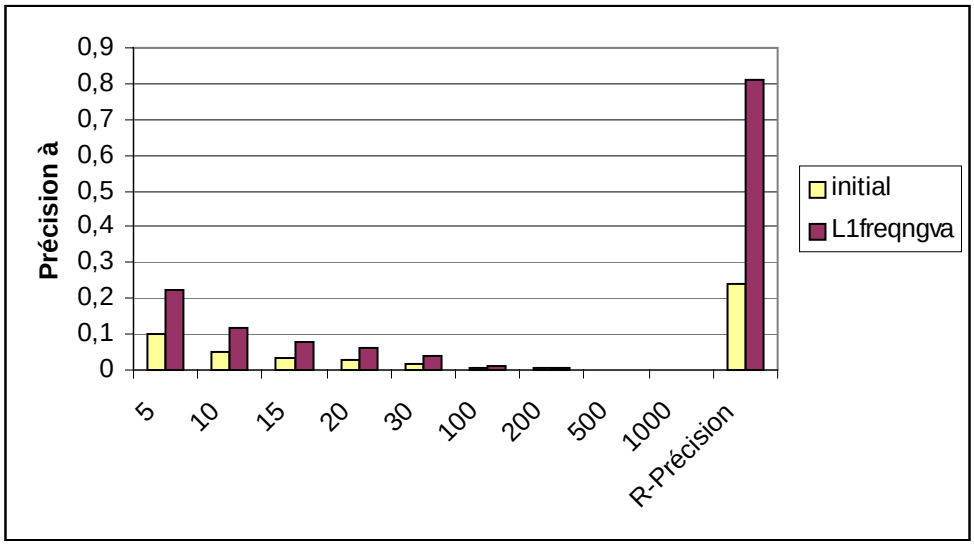


Figure 4. *Moyennes Rappel-Précision interpolées, Précision à différents niveaux et R-Précision – fonction sans et avec apprentissage – Collection de test.*

Dans la collection, environ la moitié des documents ont moins de 5 catégories. Il faut noter que lorsqu'un document est associé à une seule catégorie, la précision qui peut être obtenue par le meilleur apprentissage est de 0,2 à 5 catégories (1 catégorie sur 5), 0.1 à 10 catégories, …

Bien que nous n'ayons pas obtenu un apprentissage parfait (puisque nous n'obtenons pas une précision de 1.0 quel que soit le taux de rappel), les résultats sont stables avec une précision supérieure à 85% quel que soit le taux de rappel.

6. Conclusion

Nous avons proposé un mécanisme de catégorisation automatique de textes. Ce principe consiste à identifier les catégories dans une hiérarchie les plus adéquates compte tenu du contenu d'un texte. Un texte peut être associé à différentes catégories, rendant compte ainsi de l'aspect multi-facette des textes. Cette association est pondérée. Le calcul de pondération se base sur une méthode originale selon un principe de vote.

La catégorisation de textes permet de structurer des espaces d'information suivant des hiérarchies de catégories représentant la connaissance de domaines. Chaque texte peut être associé à différentes catégories d'une hiérarchie et également à plusieurs hiérarchies. Les hiérarchies peuvent alors servir de base à une recherche d'information. Elles permettent d'éviter les ambiguïtés de termes rencontrées en recherche par texte libre, ainsi que la récupération de textes n'appartenant pas au contexte de recherche.

Le mécanisme proposé a été évalué sur la collection Reuters-21578, une des collections les plus utilisées pour évaluer les techniques de catégorisation et de classification. Cette évaluation n'a pas permis de prendre en compte la structure hiérarchique des catégories (les catégories Reuters sont planes). En revanche, de nombreux paramètres de catégorisation ont pu être étudiés. L'évaluation des performances se base sur des mesures de rappel/précision. Nous avons obtenu une R-Précision de 0,77 sur l'ensemble de test de Reuters-21578 après apprentissage sur l'ensemble d'apprentissage. Le taux de précision est supérieur à 0,82 quel que soit le taux de rappel sur cette même collection. Nous avons montré que les résultats peuvent être améliorés en augmentant le nombre de termes associés par apprentissage à chaque catégorie. Les meilleurs résultats ont été obtenus avec 150 termes par catégorie mais nous avons également montré que, en augmentant de 50 à 150 cette taille, l'amélioration des performances n'était que de 4%.

De notre point de vue, ce principe de catégorisation ou d'indexation contrôlée par le vocabulaire issu des catégories se rapproche du projet de Web sémantique initié par le W3C (Allen,01), (Berners-Lee,01). Ce projet a pour objectif de structurer le Web actuel afin de faciliter en particulier des traitements automatiques intelligents et de permettre une coopération plus étroite entre utilisateurs et agents. Ce projet doit reposer sur les technologies XML et RDF qui permettent d'ores et déjà d'ajouter des éléments de structure aux contenus des pages. Un autre élément de base du Web sémantique correspond aux ontologies. Une ontologie, qui en philosophie réfère à la science de l'existence, correspond pour les acteurs du Web aux relations formelles entre termes. Une façon d'ajouter de la structure au Web consiste donc à créer des liens entre le contenu des pages et une ou plusieurs

ontologies, apportant ainsi une sémantique utile pour les moteurs de recherche par exemple. Notre approche se base sur des hiérarchies de catégories ou de concepts qui est une forme simple d'ontologie dans laquelle seul le lien sémantique « est-un » est représenté. Implicitement, d'autres liens sémantiques sont utilisés dans notre approche, en particulier via la radicalisation et l'apprentissage. Nos travaux futurs visent à modéliser ces liens sémantiques directement dans la structure de catégories, sous la forme d'une ontologie plus complète.

7. Références

J. Allen, « Making a semantic Web », http://www.netcrucible.com/semantic.html, 2001

J. Augé, K. Englmeier, G. Hubert, J. Mothe, « Classification automatique de textes basée sur des hiérarchies de concepts », Veille stratégique, scientifique et technologique, Barcelone, p. 291-300, 2001.

T. Berners-Lee, J. Hendler, O. Lassila, « The Semantic Web », Scientific American, 2001, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html.

S. Chaudhuri, U. Dayal, « An overview of data warehousing and OLAP technology », ACM SIGMOD Record, Vol. 26, N.1, p. 65-74, 1997.

K. K. Chin, « Support Vector Machines applied to Speech Pattern Classification », Mphil. In Computer Speech and Language Processing, Cambridge University Engineering Department, 1999.

S. Dumais, J. Platt, D. Heckerman, M. Sahami, « Inductive learning algorithms and representations for text categorization », *ACM-CIKM'98*, p. 148-155, 1998

S. Dumais, H. Chen, « Hierarchical classification of Web documents », 23nd Annual International ACM Conference on Research and Development in Information Retrieval SIGIR'2000, Athenes, 2000.

K. Englmeier, J. Mothe, « Thustworthy personal assistance: a design objective agents, Association for information systems », 7th Americas Conference on Information Systems, (Cédérom), Boston, Août 2001.

W. Frakes. « Stemming algorithms », Information Retrieval: Data Structures & Algorithms, W. Frakes and R. Baeza-Yates editors, Prentice Hall, p. 131-160, 1992.

M.A. Hearst, C. Karadi, « Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy », International Conference on Research and Development in Information Retrieval, p. 246-255, 1997.

IRAIA, Projet soutenu par la commission Européenne via le 5ième programme cadre, Getting Orientation in Complex Information Spaces as an Emergent Behavior of Autonomous Information Agents, IST-1999-10602.

T. Joachims, « Text categorization with Support Vector Machines: Learning with many relevant features », 10th European Conference on Machine Learning ECML'98, p. 137-142, 1998.

T. Joachims, « A probabilistic analysis of the Rocchio algorithm with tf.idf for text categorization », 14th Inter. Conf. On Machine Learning, ICML97, p. 143-151, 1997.

D. Koller, M. Sahami, « Hierarchically classifying documents using very few words », 14th Inter. Conf. On Machine Learning, ICML97, p. 170-178, 1997.

D. D. Lewis, M.A. Ringuette, « A comparison of two learning algorithms for text categorisation », 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR'94, p. 81-93, 1994.

D.D. Lewis, R.E. Schapire, J.P. Callan, R. Papka, « Training algorithms for linear text classifiers », 19th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR'96, p. 298-306, 1996.

Y. Li, « Toward a qualitative search engine », IEEE Internet Computing, vol. 2, n° 4, p. 24-29, 1998.

J. Mothe, « Recherche et exploration d'informations – Découvertes de connaissances pour l'accès à l'information », Habilitation à diriger des recherches, Université P. Sabatier, Décembre 2000.

J. Mothe, C. Chrisment, J. Alaux, « Visualisation globale de collections de documents sous forme d'hypercube - Le système DocCube », Journées francophones d'Extraction et de Gestion des Connaissances, EGC, 2002.

B. Pauer, P. Holger, « Statfinder », Document Package Statfinder, Vers. 1.8, mai 2000.

J.M. Ponte, W.B. Croft, « A language modeling approach to information retrieval », 21nd International Conference on Research an Development in Information Retrieval, p. 275-281, 1998.

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

J.J. Rocchio, « Relevance feedback in information retrieval », The smart Retrieval System – Experiments in Automatic Document Processing, p. 313-323, Prentice-Hall, Englewood, Cliffs, New Jersey, 1971.

G. Salton, « The SMART Retrieval System », Experiments in automatic document processing, Prentice Hall Inc., Englewood Cliffs, NL, 1971.

H. Schutze, D. Hull, J.O. Pedersen, « A comparison of classifiers and document representations for the routing problem », 18th Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR'95, p. 229-237, 1995.

K. SparkJones, C. J. Van Rijsbergen, « Progress in documentation », Journal of Documentation, 32(1), p. 59-75, 1976.

TREC, Text Retrieval Conference, http://trec.nist.gov.

C. J. Van Rijsbergen, « Information Retrieval », Butterworths, London, Second Edition, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html

V. Vapnik, « The Nature of Statistical learning Theory », Springer-Verlag, 1995.

R. Vinot, F. Yvon, « Quand simplicité rime avec efficacité : analyse d'un catégoriseur de textes », Colloque International sur la Fouille de Texte (CIFT), Hammamet, 2002

E.M. Voorhees, « Implementing agglomerative hierarchic clustering algorithms for use in document retrieval », Information Processing & Management, vol. 22 n°6, p 465-476, 1986.