# Indexation sémantique au moyen de coupes de redondance minimale dans une ontologie

Florian Seydoux & Jean-Cédric Chappelier
Faculté Informatique et Communications
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH–1015 Lausanne, Suisse
{florian.seydoux,jean-cedric.chappelier}@epfl.ch

**Mots-clefs :** Indexation sémantique, Recherche documentaire, Redondance minimale, Ontologie.

**Keywords:** Semantic Indexing, Information Retrieval, Minimal Redundancy, Ontology.

**Résumé** Plusieurs travaux antérieurs ont fait état de l'amélioration possible des performances des systèmes de recherche documentaire grace à l'utilisation d'indexation sémantique utilisant une ontologie (p.ex. WordNet). La présente contribution décrit une nouvelle méthode visant à réduire le nombre de termes d'indexation utilisés dans une indexation sémantique, en cherchant la coupe de redondance minimale dans la hiérarchie fournie par l'ontologie. Les résultats, obtenus sur diverses collections de documents en utilisant le dictionnaire EDR, sont présentés.

**Abstract** Several former works have shown that it is possible to improve information retrieval performances using semantic indexing, adding additional information coming from a thesaurus (e.g. WordNet). This paper presents a new method to reduce the number of "concepts" used to index the documents, by determining a minimum redundancy cut in the hierarchy provided by the thesaurus. The results of experiments carried out on several standard document collections using the EDR thesaurus are presented.

## 1 Introduction

L'utilisation de connaissances sémantiques dans le cadre de la Recherche Documentaire (RD) n'est pas nouvelle. On voit se dégager dans la littérature scientifique principalement trois champs d'application : *l'expansion de requêtes* (Voorhees, 1994; Moldovan & Mihalcea, 2000), *la désambiguïsation sémantique (WSD)* (Ide & Véronis, 1998; Wilks & Stevenson, 1998) et *l'indexation sémantique*. C'est dans ce dernier cadre que se situe le travail présenté ici.

L'indexation sémantique consiste à utiliser, pour indexer des documents, le(s) sens des mots qu'ils contiennent, au lieu ou en plus des mots[1] eux-mêmes comme c'est le cas en RD classique,

Ce travail a été financé par le projet n°200020–103529 du Fond National Suisse pour la Recherche Scientifique.

[1] Habituellement, leurs lemmes ou leurs racines (*stems*).

de manière à améliorer tant le rappel (par le biais des relations de synonymie) que la précision (en traitant correctement les cas d'homographie/polysémie).

Les différentes expériences rapportées à ce sujet dans la littérature font cependant état de résultats peu concluants, parfois même contradictoires : si certains observent que l'ajout de ce type d'information, réalisée de manière automatique, dégrade les performances de leur système (Salton, 1968; Harman, 1988; Voorhees, 1993; Voorhees, 1998), pour d'autres au contraire une amélioration significative est obtenue (Richardson & Smeaton, 1995; Smeaton & Quigley, 1996; Gonzalo *et al.*, 1998a; Gonzalo *et al.*, 1998b; Mihalcea & Moldovan, 2000).

Bien qu'il semble souhaitable pour un système de RD de prendre en compte un maximum d'informations, en particulier des informations de nature sémantique, un tel accroissement des termes d'indexation peut se révéler contre-productif, ou tout du moins ne pas développer son plein potentiel. En effet, une forte augmentation du nombre de termes d'indexation a non seulement comme conséquences de prolonger notablement les temps de traitement, mais surtout affecte les performances sur le plan de la précision : tenter de discriminer quelques documents parmi un ensemble sur la base d'un très grand nombre de critères est difficile à réaliser, la « distance » – généralement une similarité ou une dissemblance – entre chaque paire de documents tendant à devenir à peu près la même (effet « *curse of dimensionality* »).

Ce problème n'est pas nouveau et il existe déjà un certain nombre de techniques visant à limiter la taille du jeu d'indexation : en plus de celles procédant par filtrage (en utilisant par exemple un anti-dictionnaire (*stoplist*), la catégorie morpho-syntaxique, ou encore les fréquences d'occurrence), la limitation du nombre de termes d'indexation a aussi été envisagée au moyen de techniques statistiques issues de l'analyse des données (analyse en composantes principales, analyse factorielle discriminante) (Deerwester *et al.*, 1990; Hofmann, 1999). Cependant, la plupart de ces techniques ne sont pas nécessairement adaptées lorsque l'on est en présence d'informations supplémentaires sur les termes d'indexation ayant une structure formelle (au lieu de statistique). L'objectif des travaux présentés dans cette contribution est précisément d'utiliser une ressource sémantique externe (i.e. additionnelle aux données de recherche documentaire proprement dites) structurée, de type ontologie, en vue d'augmenter la richesse de l'indexation. La spécificité de ce travail par rapport à des travaux antérieurs similaires, qui utilisent des « *synsets* » ou des hyperonymes de *WordNet* comme termes d'indexation (Gonzalo *et al.*, 1998a; Gonzalo *et al.*, 1998b; Whaley, 1999; Mihalcea & Moldovan, 2000), est d'essayer de faire un pas supplémentaire en sélectionnant les « concepts » à utiliser comme termes d'indexation au moyen d'un critère issu de la théorie de l'information, la *Coupe de Redondance Minimale* (CRM, voir figure 1), que l'on applique à la relation inclusive « est-un » (hyperonymie) obtenue ici par le biais de la taxonomie (anglaise) *EDR* (Miyoshi *et al.*, 1996).

# 2 Coupe de redondance minimale

## 2.1 Objectifs

Le choix du « concept hyperonyme »[2] à utiliser pour représenter un mot est un choix délicat : un concept trop général dégradera les performances du système en diminuant la précision, tandis

[2] Nous désignons par « concept hyperonyme » un nœud non feuille dans l'ontologie. Les feuilles de l'ontologie représentent les mots.

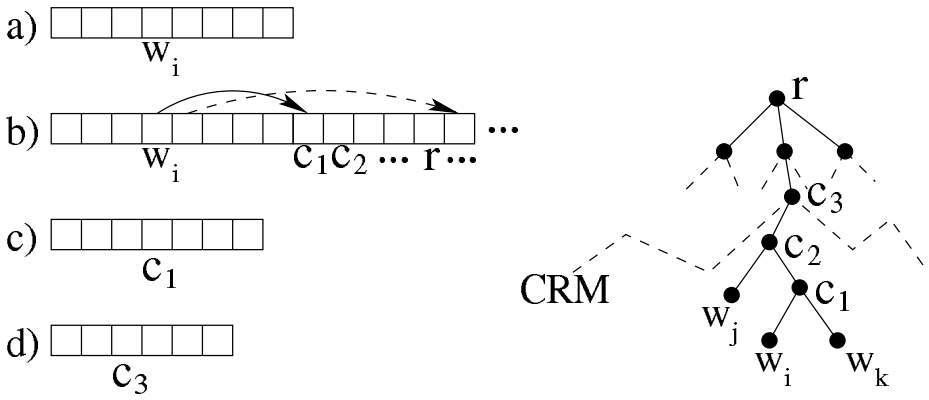


FIG. 1 – Différentes méthodes d'indexation : (a) traditionnelle, au moyen des mots, racines (*stems*) ou lemmes ; (b) utilisant une ontologie sémantique (illustration de droite), chaque terme d'indexation de (a) est augmenté par tout ou partie des « concepts » le recouvrant ; cela conduit à une explosion du nombre de termes d'indexation ; (c) indexation par les concepts de plus bas niveau ($\approx$ indexation par « synsets ») : chaque terme d'indexation est *remplacé* par son concept hyperonyme direct, factorisant ainsi tous les mots dominés par ce concept ; on réduit donc le nombre de termes d'indexation, tout en permettant de détecter la similarité entre documents contenant ces mots ; (d) indexation par une Coupe de Redondance Minimale (CRM) : chaque terme d'indexation est remplacé par l'un de ses concepts hyperonymes, déterminé par la CRM. Cela restreint d'avantage le nombre de termes d'indexation, le nombre de mots couverts (factorisés) par chacun d'eux étant plus grand qu'avec le concept hyperonyme direct.

qu'un concept trop spécifique ne permettra pas de réduire significativement le nombre de termes d'indexation et conservera la distinction entre mots de sens proches.

Pour déterminer le niveau adéquat des concepts d'indexation, nous faisons ici le choix de ne prendre en considération des coupes dans l'ontologie (une coupe étant un ensemble minimal[3] de nœuds définissant une partition sur les feuilles), en considérant que chaque nœud représente alors l'ensemble des feuilles qu'il recouvre.

Le problème est de trouver une stratégie permettant d'identifier une coupe « optimale » en un temps acceptable. Pour une tâche relativement similaire, Li (1998) propose d'utiliser le critère MDL (*Minimum Description Length*). Si ce critère est facilement calculable, il a comme inconvénient, du moins lorsque appliqué à l'ontologie EDR, de très souvent sélectionner la racine de l'ontologie comme coupe « optimale » ; ce qui n'est pas vraiment adéquat pour la tâche considérée ! Nous nous proposons donc ici d'employer un autre critère, fondé sur la théorie de l'information, permettant d'identifier une coupe pour laquelle la *redondance d'information* est minimale, c'est-à-dire une coupe qui équilibre le plus possible les degrés de description des mots factorisés en tenant compte de la probabilité d'occurrence de ces mots.

## 2.2 Critère de redondance minimale

Soient $\mathcal{N} = \{n_i\}$ l'ensemble des nœuds (concepts ou mots) et $\mathcal{W}$ l'ensemble des feuilles (mots uniquement) contenus dans l'ontologie considérée. On définit alors une coupe $\Gamma$ comme un sous-ensemble minimal[3] de $\mathcal{N}$ recouvrant $\mathcal{W}$. Une coupe probabilisée $M = (\Gamma, P)$ est une

[3] Par « minimal », on entend qu'aucun nœud de la coupe ne peut en être retiré sans en diminuer la couverture.

paire composée d'une coupe $\Gamma$ et d'une distribution de probabilités $P$ sur $\Gamma$. On notera $|\Gamma|$ le nombre de nœuds de la coupe (et par extension: $|M| = |\Gamma|$).

Dans la suite, nous considérons la coupe $M = (\Gamma, P_f)$ probabilisée par les fréquences d'occurrences des mots correspondant aux feuilles de l'ontologie : $P_f(n_i) = f(n_i)/|D|$, où $f(n_i)$ représente le nombre d'occurrences du concept (ou mot) $n_i$ dans les données $D$. Pour calculer $f(n_i)$, on admet qu'il y a occurrence de $n_i$ lorsqu'il y a occurrence de l'un des $w_i \in {n_i}^{++}$ mots hyponymes de $n_i$, où $n^{++}$ représente la fermeture transitive de $n^+$, ensemble des successeurs de $n$.

La redondance $R(M)$ d'une coupe probabilisée $M = (\Gamma, P)$ est définie par (Shannon, 1948):

$$R(M) = 1 - \frac{H(M)}{\log |M|}, \quad \text{avec} \qquad H(M) = -\sum_{n \in \Gamma} P(n) \cdot \log P(n).$$

Minimiser la redondance revient à maximiser le rapport entre l'entropie des éléments de la coupe et sa valeur maximale possible ($\log |M|$) ; le but est donc de trouver une coupe probabilisée $M$ qui maximise le critère $\mathcal{C}_H$:

$$\mathcal{C}_H = \begin{cases} 0 & \text{si } |M| \leq 1, \\ \frac{H(M)}{\log |M|} & \text{sinon.} \end{cases}$$

Un tel critère pose cependant quelques difficultés en pratique: d'une part, il ne permet pas d'identifier une coupe optimale unique, mais un *ensemble* de coupes possibles ; d'autre part, l'optimum local sur une partie de l'ontologie est conditionné par l'optimum sur le reste (et inversement). Pour identifier les modèles satisfaisant le critère global, il faudrait donc le calculer pour l'ensemble des coupes possibles.

La première difficulté peut être surmontée de manière relativement aisée, par exemple en ne retenant qu'une coupe choisie au hasard, ou en favorisant celles admettant le plus de nœuds, ou encore en guidant le choix selon la profondeur moyenne des nœuds.

Pour être calculable, la seconde difficulté implique par contre de renoncer à l'optimalité globale. Néanmoins, il est possible d'utiliser un algorithme de programmation dynamique permettant d'obtenir une coupe acceptable (heuristique). Cet algorithme consiste à choisir, pour un sous-arbre[4] dans l'ontologie, une coupe optimale parmi celles constituées des successeurs directs de la racine de ce sous-arbre et les sous-coupes « optimales » de chacun de ces successeurs, obtenues de manière similaire. Plus formellement, l'algorithme récursif donné en table 1 est appliqué à partir de la racine de l'ontologie[5].

## 2.3 Exemple

Pour illustrer le fonctionnement de la technique de sélection des coupes décrite précédemment, admettons que l'on dispose de l'ontologie présentée en figure 2 ; les valeurs indiquées en regard

[4] Bien que les ontologies utilisées présentent usuellement une structure de graphe orienté sans cycle (DAG), nous simplifierons ici le propos en considérant qu'il s'agit d'arbres. Cette approximation, qui n'invalide en rien les raisonnements exposés ici, n'est évidemment pas faite en pratique.

[5] En pratique, plusieurs optimisations sont introduites (notamment, les successeurs feuilles d'un nœud sont nécessairement compris dans la sous-coupe optimale pour ce nœud) ; mais elles ne changent rien à l'aspect fondamental présenté ici.

ALGORITHME CRM

Entrée : un nœud $t$ (dans une hiérarchie).

Sortie : $CRM$ : *une* coupe de redondance minimale sous ce nœud.

**Si** $t \in \mathcal{W}$
  $CRM \leftarrow \{t\}$
**Sinon**
  **Pour** $n_i \in t^+$
    $\gamma_i \leftarrow \text{CRM}(n_i)$
    $\vartheta_i \leftarrow \{n_i\}$
  **Pour** $1 \leq k \leq n := |t^+|$
    $\Gamma_k \leftarrow \bigcup_{j \in [1:n \setminus k]} \gamma_j \cup \vartheta_k$
  $\Gamma_{n+1} \leftarrow \bigcup_{j \in [1:n]} \gamma_j$
  $\Gamma_{n+2} \leftarrow \bigcup_{j \in [1:n]} \vartheta_j$
  $CRM \leftarrow \text{Argmax}_{\Gamma_j : 1 \leqslant j \leqslant n+2} \left( \mathcal{C}_H(\Gamma_j) \right)$

où Argmax retourne une coupe possible réalisant ce maximum.

TAB. 1 – Algorithme de recherche heuristique d'une CRM.

des feuilles correspondent aux fréquences d'occurrences des mots y-relatifs obtenues sur un corpus fictif.

Pour la coupe $\Gamma = [\text{ANIMAL}, \text{PLANTE}, \text{TRANSPORT}]$, on obtient la valeur du critère $\mathcal{C}_H$:

| $n_i$ | ANIMAL | PLANTE | TRANSPORT |
|---|---|---|---|
| $f(n_i)$ | 18 | 30 | 1 |
| $P_f(n_i)$ | 0.3673 | 0.6122 | 0.0204 |
| $-P_f(n_i)\log_2 P_f(n_i)$ | 0.5307 | 0.4334 | 0.1146 |
| $\mathcal{C}_H(\Gamma) = \frac{1.0787}{\log_2(3)} = 0.6806$ | | | |
| $R(\Gamma) = 1 - \mathcal{C}_H(\Gamma) = 0.3194$ | | | |

Dans un tel cas de figure, en examinant l'ensemble des 2036 différentes coupes possibles, on trouverait que le critère sur la coupe optimale (indiquée sur la figure 2) vaut 0.874. L'algorithme de recherche par optimum local trouve une coupe pour laquelle le critère est légèrement inférieur: 0.810 ; mais son obtention ne nécessite l'évaluation que de 36 coupes différentes.

# 3 Expériences

Nous avons effectué un jeu d'expériences en utilisant les collections standards ADI, TIME, MED, CACM et CISI[6] du projet SMART (Salton, 1971), ainsi qu'une ontologie produite à partir du dictionnaire électronique *EDR* (Miyoshi *et al.*, 1996).

*EDR* est organisée en cinq dictionnaires de différents types, plus ou moins indépendants les uns

[6] Disponibles à l'adresse *ftp://ftp.cs.cornell.edu/pub/smart/*.

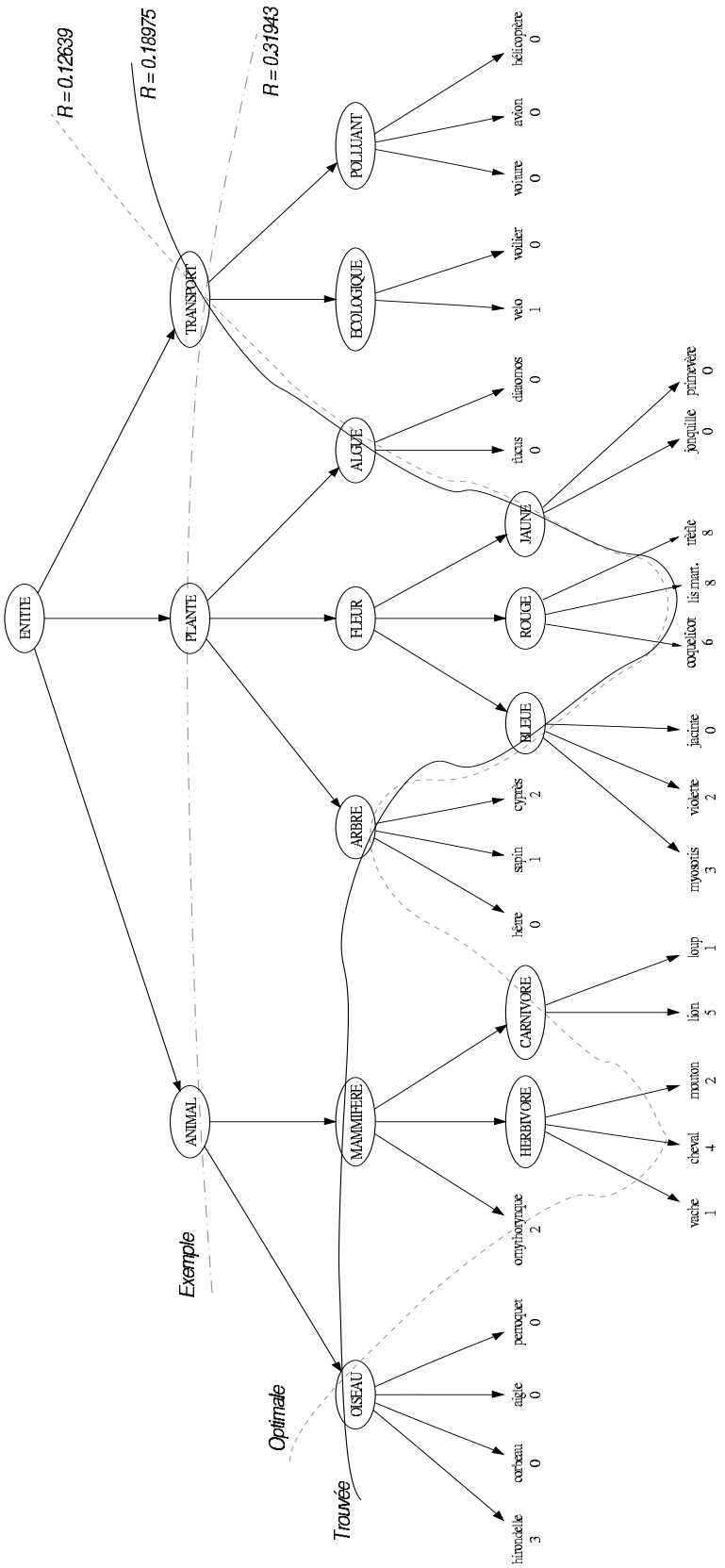


FIG. 2 – Exemple de coupes dans une ontologie.

des autres. Parmi l'ensemble de ces dictionnaires, les deux suivants sont utilisés pour constituer l'ontologie:

**le dictionnaire des mots anglais,** qui rassemble les informations morphologiques (prononciation, découpage syllabique, inflexion, ...) et syntaxiques (catégorie morpho-syntaxique, dénombrabilité, flexions, ...) pour un peu plus de 240′000 graphies différentes (correspondant à ≈ 420′000 mots), et permet de relier ces graphies avec les informations du dictionnaire des concepts. Les graphies de ce dictionnaire sont principalement (mais pas exclusivement) des lemmes ; il comporte également un nombre important de multi-termes (> 113′000), figurant des mots composés et expressions idiomatiques.

**le dictionnaire des concepts,** qui décrit à peu près 490′000 concepts, organisés hiérarchiquement entre eux selon des relations d'hyponymie/hyperonymie (chaque concept pouvant avoir plusieurs hyponymes et hyperonymes). Un certain nombre de relations sémantiques binaires supplémentaires (telles que objet-action, agent-action, agent-but) sont par ailleurs décrites, mais nous ne les utilisons pas ici. Remarquons qu'un nombre important de concepts (environ la moitié) ne sont pas directement associés à des mots ; ces concepts ne peuvent être définis et appréhendés qu'au travers de leurs relations avec les autres concepts.

Le système de RD utilisé est le modèle vectoriel SMART, combiné à un lemmatiseur externe[7], qui fait également office de segmenteur (*tokenizer*) et d'étiqueteur morpho-syntaxique. Un filtrage par catégorie grammaticale est réalisés (ne sont conservés que les noms, adjectifs et verbes), mais nous n'utilisons pas d'anti-dictionnaire et ne faisons pas de filtrage fréquentiel.

Les transformations du jeu d'indexation sont obtenues en prétraitant les données soumises au système de RD:

1. en premier lieu, les diverses informations textuelles (principalement titre et contenu) des documents sont agrégées, et les autres informations (auteurs, sources, etc.) supprimées ; documents et requêtes sont ensuite segmentés et lemmatisés ;
2. on cherche ensuite les correspondances entre les mots contenus dans les documents et ceux décrits dans l'ontologie ; on tente d'établir en priorité une correspondance avec la graphie, et s'il n'y en a pas, avec sa forme lemmatisée ; les mots sans correspondance sont indexés de manière traditionnelle ; les taux de couverture[8] sur les différentes collections sont de l'ordre de 90%.
3. on procède ensuite à l'expansion de la hiérarchie des concepts relatifs aux mots conservés pour l'ensemble des documents ; selon les différents cas expérimentés, on prendra soit la *totalité des concepts* possibles (en tablant sur un renforcement mutuel des concepts « corrects » induit par les multiples co-occurrences), soit uniquement le *concept le plus probable* (dans l'absolu pour le mot donné – cette information est présente dans l'ontologie utilisée) ;
4. on détermine ensuite une coupe optimale selon le critère $C_H$, au moyen de l'algorithme CRM présenté en section 2.2 ;
5. finalement, on substitue les mots des documents et des requêtes par les identificateurs des concepts de la coupe qui les subordonnent.

[7] Le système Sylex 1.7 (© 1993-98 DECAN INGENIA).

[8] Par « *couverture* », on désigne la fraction des occurrences des mots couverts par l'ontologie.

| | *mesure* | (a) | (b) | (c) | (d) |
|---|---|---|---|---|---|
| corpus ADI (82 documents) | | | | | |
| tous les concepts, tf.idf | taille index | 1800 | 14748 | 10099 | 1292 |
| | précision | **0.3578** | 0.3134 | 0.3356 | 0.2458 |
| | rappel | 0.6984 | 0.7126 | **0.7406** | 0.6017 |
| tous les concepts, sans pondération | précision | 0.2497 | 0.1219 | **0.2550** | 0.1607 |
| | rappel | 0.5996 | 0.3452 | **0.6708** | 0.5130 |
| concept le plus probable, tf.idf | taille index | 1800 | 5255 | 2888 | 658 |
| | précision | 0.3578 | 0.4060 | **0.4274** | 0.2052 |
| | rappel | 0.6984 | 0.7306 | **0.7217** | 0.5200 |
| concept + probable, sans pondération | précision | 0.2497 | 0.1376 | **0.2939** | 0.1466 |
| | rappel | 0.5996 | 0.3727 | **0.7141** | 0.4911 |
| corpus TIME (423 documents) | | | | | |
| tous les concepts, tf.idf | taille index | 21815 | 93707 | 70091 | 6760 |
| | précision | **0.5496** | 0.4231 | 0.4536 | 0.2683 |
| | rappel | **0.8901** | 0.7642 | 0.8036 | 0.6026 |
| tous les concepts, sans pondération | précision | **0.3288** | 0.0337 | 0.2353 | 0.0370 |
| | rappel | **0.7755** | 0.1021 | 0.5709 | 0.1387 |
| concept le plus probable, tf.idf | taille index | 21815 | 53140 | 31612 | 4814 |
| | précision | 0.5496 | 0.5143 | **0.5565** | 0.2729 |
| | rappel | 0.8901 | 0.8760 | **0.9053** | 0.5162 |
| concept + probable, sans pondération | précision | 0.3288 | 0.0346 | **0.3692** | 0.0372 |
| | rappel | **0.7755** | 0.1201 | 0.7590 | 0.1322 |
| corpus MED (1033 documents) | | | | | |
| tous les concepts, tf.idf | taille index | 11893 | 51712 | 38524 | 4078 |
| | précision | **0.4607** | 0.3029 | 0.2996 | 0.2336 |
| | rappel | **0.5547** | 0.3903 | 0.3794 | 0.3142 |
| tous les concepts, sans pondération | précision | **0.3623** | 0.0105 | 0.1905 | 0.0229 |
| | rappel | **0.4574** | 0.0246 | 0.2749 | 0.0513 |
| concept le plus probable, tf.idf | taille index | 11893 | 30284 | 18109 | 2888 |
| | précision | **0.4607** | 0.4266 | 0.4518 | 0.0743 |
| | rappel | **0.5547** | 0.5169 | 0.5404 | 0.1042 |
| concept + probable, sans pondération | précision | **0.3623** | 0.0105 | 0.3229 | 0.0132 |
| | rappel | **0.4574** | 0.0313 | 0.4230 | 0.0368 |
| corpus CISI (1460 documents) | | | | | |
| tous les concepts, tf.idf | taille index | 10019 | 53453 | 39544 | 3516 |
| | précision | **0.1733** | 0.1043 | 0.1139 | 0.0740 |
| | rappel | **0.2318** | 0.1627 | 0.1675 | 0.1294 |
| tous les concepts, sans pondération | précision | **0.0687** | 0.0232 | 0.0569 | 0.0282 |
| | rappel | **0.1239** | 0.0376 | 0.0963 | 0.0492 |
| concept le plus probable, tf.idf | taille index | 10019 | 26246 | 14993 | 1894 |
| | précision | 0.1733 | 0.1590 | **0.1825** | 0.0602 |
| | rappel | **0.2318** | 0.2131 | **0.2313** | 0.0895 |
| concept + probable, sans pondération | précision | 0.0687 | 0.0201 | **0.0805** | 0.0221 |
| | rappel | 0.1239 | 0.0403 | **0.1300** | 0.0435 |
| corpus CACM (3204 documents) | | | | | |
| tous les concepts, tf.idf | taille index | 10053 | 51712 | 38524 | 4078 |
| | précision | **0.2865** | 0.1293 | 0.1935 | 0.1089 |
| | rappel | **0.4534** | 0.2579 | 0.3617 | 0.1999 |
| tous les concepts, sans pondération | précision | **0.1555** | 0.0133 | 0.1447 | 0.0320 |
| | rappel | **0.3082** | 0.0306 | 0.2549 | 0.0699 |
| concept le plus probable, tf.idf | taille index | 10053 | 25207 | 14681 | 2670 |
| | précision | **0.2865** | 0.2358 | 0.2804 | 0.0645 |
| | rappel | 0.4534 | 0.3834 | **0.4567** | 0.1090 |
| concept + probable, sans pondération | précision | **0.1555** | 0.0230 | 0.1472 | 0.0245 |
| | rappel | **0.3082** | 0.0302 | 0.2926 | 0.0385 |

TAB. 2 – Résultats des différentes expériences sur différents corpus. (a) : mots uniquement ; (b) : mots + concepts ; (c) : hyperonymes directs et (d) : hyperonymes dans CRM (cf aussi fig. 1).

On trouvera dans la table 2 les valeurs de précision (« *11-pt prec* ») et de rappel (« *30 doc* »)[9] fournies par le système SMART. Toutes les expériences sont par ailleurs conduites en utilisant soit le schéma de pondération classique (« tf.idf »), soit sans pondération.

On constate que l'indexation par hyperonymes directs obtient des résultats sensiblement égaux au système de base, mais pour un rappel plus élevé. L'indexation par CRM dégrade par contre les performances.

# 4 Conclusion

Les résultats obtenus sur ces expériences ne sont malheureusement pas concluants quant à l'utilisation du critère CRM pour l'indexation sémantique. Cependant, plusieurs remarques sont à apporter :

– Le critère utilisé ici ne permet pas de sélectionner, ni même d'influencer, le niveau de profondeur dans l'ontologie de la coupe obtenue. Au vu de la réduction drastique du jeu d'indexation et des mauvaises performances obtenues, il semble que ce critère, ou du moins l'heuristique implémentée, sélectionne une coupe située trop haut dans la hiérarchie, ce qui a comme conséquence évidente de faire baisser la précision. La bonne performance de la coupe au niveau des concept hyperonymes directs nous permet de croire qu'il doit y avoir un niveau plus adapté, plus proche des feuilles, pour la CRM.
  On pourrait par exemple limiter considérablement l'espace de recherche de la coupe idéale en empêchant de considérer des nœuds situés « trop hauts » dans la hiérarchie. Une piste à explorer pour améliorer tant l'adéquation de la coupe sélectionnée avec un processus d'indexation que la recherche de cette coupe elle-même consisterait à explorer les gains possibles en terme de redondance à partir de la coupe uniquement constituée de feuilles, et en dirigeant la recherche vers le haut de la hiérarchie, plutôt que de haut en bas à partir de la racine, comme dans l'heuristique présentée ici.
– Par ailleurs, en conservant l'idée d'une action sur le jeu d'indexation lui-même, il serait intéressant d'examiner de quelle manière les pondérations (e.g. « tf.idf »), utilisées uniquement lors de la recherche des documents proprement dite, devraient être prises en compte lors de la détermination de la coupe.
– Finalement, les résultats présentés ici restent à corroborer avec ceux à obtenir avec d'autres ontologies, en particulier WordNet, qui a une structure assez différente d'EDR.

Pour terminer, soulignons que l'intérêt de la technique présentée dépasse le cadre de la stricte recherche documentaire. Celle-ci pourrait en effet s'avérer utile, et peut être même plus prometteuse, pour d'autres domaines d'application tels que la classification de documents ou le résumé automatique.

# Références

DEERWESTER S. C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W. & HARSHMAN R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6),

[9] Il s'agit là de mesures standard: la « *11-pt precision* » est la moyenne des précisions pour les taux de rappels 0.0, 0.1, ..., 1.0, où la précision au taux de rappel 0.0 est la précision maximale obtenue sur l'ensemble des documents pertinents retrouvés ; le « *rappel 30 doc* » est le taux de rappel après 30 documents retournés.

391–407.

GONZALO J., VERDEJO F., CHUGUR I. & CIGARRAN J. (1998a). Indexing with WordNet synsets can improve text retrieval. In *Proc. of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing*, p. 38–44.

GONZALO J., VERDEJO F., PETERS C. & CALZOLARI N. (1998b). Applying EuroWordNet to multilingual text retrieval. *Journal of Computers and the Humanities*, **32**(2-3), 185–207.

HARMAN D. (1988). Towards interactive query expansion. In *Proc. of the 11th Annual Int. ACM-SIGIR Conference on Research and development in information retrieval*, p. 321–331.

HOFMANN T. (1999). Probabilistic latent semantic indexing. In *proc. of the 22th International Conference on Research and Development in Information Retrieval (SIGIR)*, p. 50–57.

IDE N. & VÉRONIS J. (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, **24**(1), 1–40.

LI H. (1998). A probabilistic approach to lexical semantic knowledge acquisition and structural disambiguation. Master's thesis, Graduate School of Science, University of Tokyo.

MIHALCEA R. & MOLDOVAN D. (2000). Semantic indexing using WordNet senses. In *Proc. of ACL Workshop on IR & NLP*.

MIYOSHI H., AMD M. KOBAYASHI K. S. & OGINO T. (1996). An overview of the EDR electronic dictionary and the current status of its utilization. In *Proc. of COLING*, p. 1090–1093.

MOLDOVAN D. I. & MIHALCEA R. (2000). Using wordnet and lexical operators to improve internet searches. *IEEE Internet Computing*, **4**(1), 34–43.

RICHARDSON R. & SMEATON A. F. (1995). *Using WordNet in a Knowledge-Based Approach to Information Retrieval*. Rapport interne CA-0395, Dublin City University, Glasnevin, Dublin 9, Ireland.

SALTON G. (1968). *Automatic Information Organization and Retrieval*. McGraw-Hill.

SALTON G. (1971). *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

SMEATON A. F. & QUIGLEY I. (1996). Experiments on using semantic distances between words in image caption retrieval. In *Proc. of 19th Int. Conf. on Research and Development in Information Retrieval*, p. 174–180.

VOORHEES E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proc. of 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, p. 171–80.

VOORHEES E. M. (1994). Query expansion using lexical-semantic relations. In *Proc. 17th Annual Int. ACM-SIGIR Conf. on Research and Development in Information Retrieval*, p. 61–69.

VOORHEES E. M. (1998). Using WordNet for text retrieval. In C. FELLBAUM, Ed., *WordNet: An Electronic Lexical Database*, chapter 12, p. 285–303. MIT Press.

WHALEY J. M. (1999). *An Application of Word Sense Disambiguation to Information Retrieval*. Rapport interne PCS-TR99-352, Dartmouth College, Computer Science, Hanover, NH.

WILKS Y. & STEVENSON M. (1998). Word sense disambiguation using optimised combinations of knowledge sources. In *Proc. of the 17th Int. Conf. on Computational Linguistics*, p. 1398–1402.