

# Mesure de similarité sémantique pour l'indexation de documents semi-structurés

Haïfa Zargayouna<sup>1</sup> et Sylvie Salotti<sup>2</sup>

<sup>1</sup>LIMSI/CNRS, Université Paris 11  
haifa.zargayouna@limsi.fr

<sup>2</sup>LIPN - CNRS UMR 7030, Université Paris 13  
sylvie.salotti@lipn.univ-paris13.fr

**Résumé** : Nous présentons dans cet article une mesure de similarité entre les concepts d'une ontologie que nous utilisons dans un système d'indexation de documents XML. Les documents sont structurés par un ensemble de balises sémantiquement pertinentes reliées à l'ontologie. Une partie des termes du corpus est également reliée à l'ontologie. Nous avons étendu le modèle vectoriel de Salton pour prendre en compte la structure des documents et le voisinage sémantique des termes. Ce système d'indexation pourrait être très utile dans le cadre d'un système de Raisonnement à Partir de Cas (RàPC) où les cas seraient décrits sous forme textuelle avec une certaine structure.

**Mots-clés** : Similarité, ontologie, sémantique, index.

## 1 Introduction

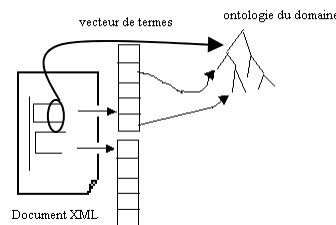
Dans certains systèmes de Raisonnement à Partir de Cas (RàPC), les cas sont décrits sous forme textuelle avec une certaine structure et il est primordial de pouvoir indexer ces cas selon leur contenu et raisonner dessus par similarité. (Lenz, 1998) compare les caractéristiques de ces systèmes de RàPC textuel à celles des systèmes de Recherche d'Information.

Nous nous intéressons essentiellement à la **phase de remémoration** des cas dans le cadre d'applications où le RàPC est utilisé pour des tâches d'aide au diagnostic ou à l'interprétation. Par exemple, à partir de la description d'un incident sur le réseau téléphonique, on recherche dans une base de fiches de descriptions d'incidents antérieurs les cas les plus similaires pour les présenter à l'opérateur afin de l'aider à définir les actions à entreprendre. Autre exemple : à partir d'un ensemble de compte-rendus d'hospitalisation on recherche ceux qui correspondent le mieux au cas d'un nouveau patient pour aider un médecin dans la prescription d'examens ou de traitements. La structure d'un compte-rendu en différentes rubriques (informations sur le patient, antécédents, symptômes...) peut être traduite à l'aide de balises dans un document XML. Nous présentons dans cet article une méthode d'évaluation de similarité mise en œuvre dans un système d'indexation sémantique de documents XML (documents textuels semi-structurés). L'avantage de ces documents est qu'ils possèdent une structure qui facilite leur présentation, ainsi que leur interprétation et leur exploitation dans des contextes présentant différents besoins. Cependant, très souvent, la majeure partie de l'information reste contenue dans les champs textuels, l'utilisation exclusive de la structure n'est donc pas suffisante. Nous proposons un système d'indexation permettant d'exploiter à la fois la structure et le contenu textuel des documents.

XML s'est imposé comme format standard de documents et un nombre de plus en plus important de documents sont disponibles en format XML. Cependant l'information apportée par les balises peut varier d'un simple découpage de la structure du document (titre, sections, paragraphe) à un véritable découpage sémantique dans lequel les balises donnent des informations sur le contenu des éléments textuels. De plus en plus de travaux visent à obtenir un tel balisage sémantique des documents, nous nous plaçons donc dans cette hypothèse. Par exemple un

ensemble de compte-rendus médicaux pourront être structurés à l'aide de balises <info-patient>, <antécédent>, <traitement>... Cette structure nous permet de considérer le document comme un ensemble d'**unités sémantiques** représentant chacune un **contexte** particulier d'occurrence des termes. Nous avons étendu le modèle vectoriel de Salton (Salton, 71) en effectuant le calcul du poids des termes pour chaque unité sémantique. Un document n'est donc plus représenté par un vecteur mais par un ensemble de vecteurs, chacun correspondant à une unité sémantique.

Par ailleurs, nous proposons d'utiliser une ontologie pour enrichir le calcul du poids des termes en intégrant la notion de voisinage sémantique. Nous supposons donc que les balises pertinentes et une partie des termes du corpus sont reliés à des concepts organisés dans une ontologie ou une taxonomie (dans un premier temps, nous n'avons utilisé que les liens de spécialisation / généralisation). Une mesure de similarité entre les concepts nous permet alors d'intégrer la notion de voisinage sémantique lors du calcul du poids des termes. Outre la représentation d'un contexte, les unités sémantiques servent à limiter l'étendue des calculs de similarité lors de la phase d'indexation. La figure 1 présente la structure de l'index.



**Fig. 1** – Structure de l'index

Dans la section suivante, nous présentons brièvement la problématique de l'utilisation d'ontologies, ou plus généralement de ressources sémantiques, dans les systèmes de recherche d'information. Nous décrivons ensuite en section 3 différentes mesures qui ont été définies pour évaluer la similarité entre les concepts d'une ontologie. Nous terminons cette section en définissant la mesure que nous proposons d'utiliser. En section 4, nous présentons comment cette similarité entre concepts est utilisée dans l'évaluation de la similarité entre documents. Nous concluons en soulignant les avantages et les limites de notre approche et en discutant le problème de la validation d'un tel système d'indexation.

## 2 Apport de la sémantique en Recherche d'Information

Les ressources sémantiques (thésaurus, ontologies, etc.) ont un apport considérable pour le traitement des documents textuels ou multimédia. Leur utilisation en Recherche d'Information (RI) peut intervenir lors de la phase de recherche ou lors de la phase d'indexation. La phase de recherche consiste à retrouver les documents les plus pertinents par rapport à une requête donnée. En général les documents retournés sont ordonnés à l'aide d'une mesure de similarité calculée entre le document et la requête. La phase d'indexation consiste à construire au préalable une structure d'accès aux documents qui facilitera la phase de recherche. Plus la phase d'indexation est sophistiquée, plus la phase de recherche sera facile.

### 2.1 Phase de recherche

L'intérêt d'utiliser des ressources sémantiques en recherche d'information est de pouvoir retourner, lors d'une recherche par similarité, les documents qui partagent avec la requête le

maximum de concepts plutôt que le maximum de mots-clés. Les réseaux sémantiques ont montré leur apport en expansion de requêtes (Lu & Keefer, 1994). Le but de l'expansion de requête est soit d'élargir l'ensemble de documents retournés ou d'augmenter la précision. Dans le premier cas, la requête peut être étendue en ajoutant des termes similaires à ceux de la requête. Dans le deuxième cas, les termes peuvent être complètement changés pour reformuler la requête, une technique utilisée dans les retours arrière sur pertinence (Buckley et al., 1994).

## 2.2 Phase d'indexation

Les documents peuvent être indexés par un groupe de concepts, où on sait qu'un tel document traite des concepts A et B mais où on ne connaît pas les relations entre eux dans le texte. Une autre méthode attribue à chaque document une description sémantique où les concepts sont représentés avec leurs relations sémantiques (Alhulou, Napoli & Nauer 2003). Cette représentation confère un grand pouvoir d'expression mais peut par ce fait ralentir les traitements et la construction des descriptions sémantiques associées à chaque document n'est pas une tâche facile.

L'indexation automatique dans les deux cas pose des problèmes notamment celui de l'ambiguïté des termes (homonymie et polysémie) et on a généralement recours à des outils sophistiqués de Traitement Automatique des Langues (TAL). Mais ces techniques ne résolvent pas totalement le problème et il faut toujours faire le compromis entre la finesse des traitements et la complexité des systèmes. (Krovetz, 1997) a montré la nécessité d'indexer par les concepts (i.e. sens des mots) ainsi que les mots. Indexer les documents par les concepts uniquement peut induire en erreur car les techniques de désambiguïsation ne sont pas complètement fiables et se baser uniquement dessus risque d'entraîner une perte d'information. Nous indexons dans notre système les termes, indépendamment du fait qu'ils soient reliés ou pas à une ontologie. Les liens sémantiques constituent ainsi un plus, mais un terme qui n'est pas relié à l'ontologie peut aussi être retrouvé. Nous montrerons aussi que le problème d'ambiguïté est pris en charge en intégrant la notion de contexte dans nos calculs de similarité.

## 3 Mesures de similarité

La détermination du degré de similarité entre deux concepts reliés à des termes d'un document est un problème qui se pose dans beaucoup d'applications : désambiguïsation, résumé automatique, extraction d'information, indexation automatique, etc.

Rada et al. (Rada et al., 1989) ont suggéré que la similarité dans un réseau sémantique peut être calculée en se basant sur les liens taxonomiques « is-a ». Plus généralement, le calcul de similarité entre concepts peut être basée sur les liens hiérarchiques de spécialisation/généralisation. Un moyen des plus évidents pour évaluer la similarité sémantique dans une taxonomie est de calculer la distance entre les concepts par le chemin le plus court.

Nous présentons dans ce qui suit quelques mesures de similarité conceptuelle. Un état de l'art complet est présenté par (Patwardham, 2003) où ces différentes mesures sont comparés par rapport à des évaluations faites par des sujets humains. Les deux premières mesures sont fondées sur la notion de contenu informationnel, que nous expliquerons plus en détail dans la section 3.1. Ces mesures utilisent WordNet (Fellbaum, 1998). WordNet peut être considéré comme un réseau sémantique où chaque nœud représente un concept du monde réel (qui peut être une entité, un artefact, un objet, etc.). Chaque nœud est composé d'un ensemble de synonymes qui représentent le même concept, cet ensemble s'appelle *synset*. Les synsets sont reliés par des arcs qui décrivent les relations entre les différents concepts. Ils sont divisés en 4 catégories (noms, verbes, adjectifs et adverbes). La relation « is-a » est restreinte aux noms et verbes.

### 3.1 La mesure de Resnik

La notion de contenu informationnel (CI) a été la première fois introduite par (Resnik, 1995). Elle utilise conjointement l'ontologie et le corpus. Le contenu informationnel d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de sa spécificité ou généralité. On dit qu'un concept général *subsume* un concept plus spécifique. La fréquence de concepts dans le corpus est calculée pour retrouver le contenu informationnel. Cette fréquence regroupe la fréquence d'apparition du concept lui-même ainsi que des concepts qu'il subsume. La formule est la suivante :

$$CI(c) = -\log(P(c)) \quad (1)$$

Où  $P(c)$  est la probabilité de retrouver une instance du concept  $c$ . Ces probabilités sont calculés par :  $frequency(c)/N$  où  $N$  est le nombre total de concepts. Voici un extrait de WordNet, le nombre attaché à chaque noeud est  $P(c)$  (Lin, 1998).

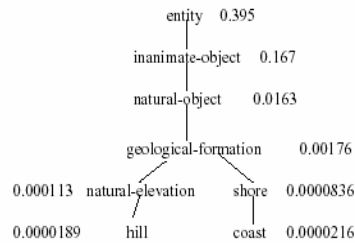


Fig. 2 – Extrait de Wordnet

Resnik définit la similarité sémantique entre deux concepts par la quantité d'information qu'ils partagent. Cette information partagée est égale au contenu informationnel du plus petit généralisant (PPG) – le concept le plus spécifique qui subsume les deux concepts dans l'ontologie.

$$Sim(c1, c2) = CI(ppg(c1, c2)) \quad (2)$$

Cette mesure ne dépend que du PPG et est de ce fait un peu sommaire car nous pouvons avoir  $ppg(a,b) = ppg(d,e)$  même si  $d$  et  $e$  sont plus proches du PPG que  $a$  et  $b$ .

### 3.2 La mesure de Jiang-Conrath

La mesure de (Jiang & Conrath, 1997) pallie aux limites de la mesure de Resnik en combinant le contenu informationnel du PPG à ceux des concepts. Elle prend en considération aussi le nombre d'arcs. Ainsi une distance est définie :

$$distance(c1, c2) = CI(c1) + CI(c2) - (2 \cdot CI(ppg(c1, c2))) \quad (3)$$

La mesure de similarité devient donc :

$$Sim(c1, c2) = 1/distance(c1, c2) \quad (4)$$

### 3.3 La mesure de Hirst-St.Onge

La mesure de (Hirst & St Onge, 1998) prend en considération toutes les relations dans WordNet. Les liens sont classés comme *haut* (eg. partie-de), *bas* (eg. sous-classe), *horizontal* (eg. antonyme). La similarité est calculée entre mots par le poids du chemin le plus court qui mène d'un terme à un autre. Il est calculé en fonctions de ces classifications qui indiquent les changements de direction :

$$\text{Sim}(c1, c2) = T - \text{chemin} - K \times d \quad (5)$$

Tels que  $T$  et  $K$  sont des constantes, *chemin* est la longueur du chemin le plus court en nombre d'arcs et  $d$  est le nombre de changements de direction.

L'idée est que deux concepts sont proches sémantiquement si leurs synsets sont connectés par un chemin qui n'est pas très long et qui ne change pas souvent de direction. S'il n'y a pas de chemin, le poids est égal à zéro.

### 3.4 La mesure de Wu-Palmer

Dans un domaine de concepts, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine. La similarité entre  $C1$  et  $C2$  est :

$$\text{ConSim}(C1, C2) = \frac{2 * N3}{N1 + N2 + 2 * N3} \quad (6)$$

Plus formellement cette mesure devient :

$$\text{ConSim}(C1, C2) = \frac{2 * \text{depth}(C)}{\text{depth}_c(C1) + \text{depth}_c(C2)} \quad (7)$$

Où  $C$  est le PPG de  $C1$  et  $C2$  (en nombre d'arcs),  $\text{depth}(C)$  est le nombre d'arcs qui sépare  $C$  de la racine et  $\text{depth}_c(C_i)$  avec  $i$  le nombre d'arcs qui séparent  $C_i$  de la racine en passant par  $C$ .

Cette mesure a l'avantage d'être simple à implémenter et d'avoir d'aussi bonnes performances que les autres mesures de similarité (Lin, 1998).

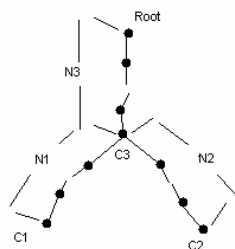


Fig. 3 – Les relations conceptuelles (Wu & Palmer, 1994)

### 3.5 Notre mesure

Dans un précédent travail (Zargayouna, 2001) dans le cadre d'une RI multimédia, nous avons calculé les similarités entre des cas formalisés en logique de description. La similarité entre deux concepts est le PPG les subsumant. Une des limites de ce travail est le manque de relation d'ordre total entre les similarités. Ce problème peut être résolu par les mesures numériques de calcul de la similarité conceptuelle. Nous l'appliquons aux données textuelles. Nous nous inspirons de la mesure de (Wu & Palmer, 1994) présentée ci-dessus. Nous n'utilisons pas la notion de contenu informationnel car elle serait redondante puisque nous combinons la mesure de similarité à la mesure distributionnelle des termes dans les documents. La mesure de (Wu & Palmer, 1994) a été utilisé par (Halkidi et al., 2003) pour organiser des documents web dans des clusters. Elle a aussi servi dans (Desmontils & Jacquin, 2001) pour évaluer la proximité sémantique de deux concepts d'une page html relativement à un thésaurus dans le cadre d'une indexation d'un site web par des ontologies.

La mesure de (Wu & Palmer, 1994) est intéressante mais présente une limite car elle vise essentiellement à détecter la similarité entre deux concepts par rapport à leur distance de leur PPG. Plus ce subsumant est général, moins ils sont similaires (et inversement). Cependant, elle ne capte pas les mêmes similarités que la similarité conceptuelle symbolique. Ainsi on peut avoir  $conSim(A, f) < conSim(A, B)$ ,  $f$  étant un des fils de  $A$  et  $B$  un des frères de  $A$ . Ce qui est à notre sens inadéquat dans le cadre de recherche d'information où il faut ramener tous les fils d'un concept (i.e requête) avant son voisinage.

Nous définissons  $spec(C1, C2)$  une fonction qui calcule la spécificité de deux concepts par rapport au concept le plus bas de l'ontologie (bottom) comme le montre la figure 3. Cette fonction servira à pénaliser les concepts qui ne sont pas dans la même lignée. Ainsi on s'assure que les fils sont pris en compte en priorité et qu'aucun concept du voisinage ne sera plus similaire que les fils.

$spec(C1, C2) = N4 * N1 * N2$  (voir figure 4) .Plus formellement :

$$spec(C1, C2) = depth_b(C) * distance(C, C_1) * distance(C, C_2) \quad (8)$$

avec  $depth_b(C)$  est le nombre maximum d'arcs qui séparent  $C$  de *bottom* et  $distance(C, C_i)$  la distance en nombre d'arcs entre  $C$  et  $C_i$

$spec(C1, C2)$  est nulle si  $C1$  est ancêtre de  $C2$  ou l'inverse. Seront pénalisés donc les concepts voisins de  $C1$  ou  $C2$ .

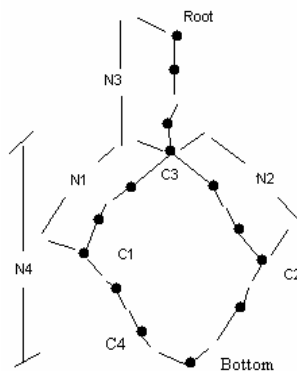


Fig. 4 – Les nouvelles relations conceptuelles

Ainsi la mesure de similarité (équation (7)) devient :

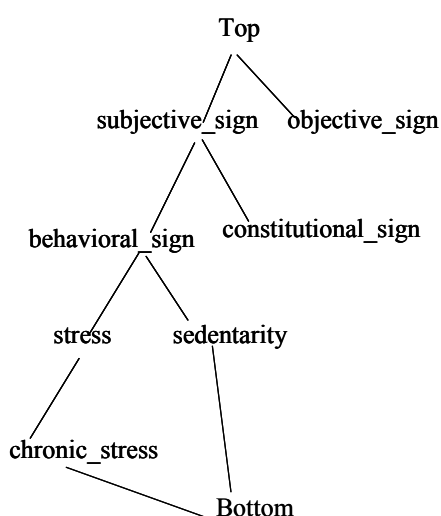
$$sim(C1, C2) = \frac{2 * depth(C)}{depth_C(C1) + depth_C(C2) + spec(C1, C2)} \quad (9)$$

Cette mesure vérifie bien les propriétés suivantes que doit vérifier une mesure de similarité :

$$S(x,y) = 1 \text{ si et seulement si } x=y$$

$$S(x,y) = S(y,x),$$

Dans l'exemple suivant, nous présentons un extrait de l'ontologie Menelas ainsi que les calculs de similarités (ConSim pour la mesure de (Wu & Palmer, 1994) et sim pour notre mesure). La similarité entre behavioral\_sign et constitutional\_sign (lien entre frères) se trouve réduite par notre mesure, celle de behavioral\_sign et chronic\_stress (lien père/fils) reste inchangée. Nous nous assurons en calculant la distance par rapport à bottom que  $\text{sim}(\text{behavioral\_sign}, \text{constitutional\_sign}) > \text{sim}(\text{behavioral\_sign}, F)$ , tel que  $F \in$  ensemble des fils de behavioral\_sign.



$$\text{ConSim}(\text{behavioral\_sign}, \text{constitutional\_sign}) = 2*1/(2+1+1) = 0.5$$

$$\text{ConSim}(\text{behavioral\_sign}, \text{chronic\_stress}) = 2*1/(2+2+0) = 0.5$$

$$\text{Sim}(\text{behavioral\_sign}, \text{constitutional\_sign}) = 2*1/(2+1+1) + (4*1*1) = 0.25$$

$$\text{Sim}(\text{behavioral\_sign}, \text{chronic\_stress}) = 2*1/(2+2+0) + (3*0*2) = 0.5$$

## 4 Similarité entre documents

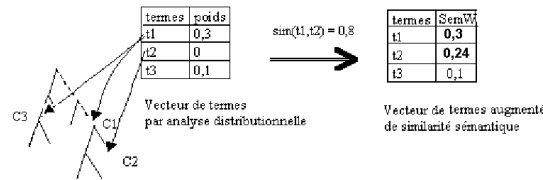
Les documents sont représentés par des ensembles de vecteurs de termes. Chaque unité sémantique génère un vecteur. Les poids des termes sont calculés en fonction de leur distribution dans les balises. Le poids d'un terme est enrichi par les similarités conceptuelles des termes co-occurents dans la même balise. Il est calculé pour un document et une unité sémantique (à savoir la balise) donnés.

Ce poids noté  $\text{SemW}(t,b,d)$  est calculé de la manière suivante :

$$\text{SemW}(t,b,d) = \text{TF-ITDF}(t,b,d) + \left( \sum_{i \in n} \text{Sim}(t,t_i) * \text{TF-ITDF}(t_i,b,d) \right) / n \quad (10)$$

avec  $\text{Sim}(t,t_i) > \text{seuil}$  ;  $t_i \in$  ensemble des  $n$  termes dans la balise  $b$  et  $\text{seuil}$  une valeur qui fixe la similarité à un certain voisinage, nous la fixons dans un premier temps à la similarité entre le concept de  $t$  et le **concept contexte** (concept qui représente la balise). TF-ITDF (Term Frequency–

Inverse Tag and Document Frequency) est le poids initial attribué aux termes en fonction du document et de la balise dans lesquels ils apparaissent (Zargayouna 2004).



**Fig. 4** – Prise en compte de la similarité sémantique

Le calcul de la similarité entre les termes co-occurents dans la même balise nous permet de gérer en partie le problème d’ambiguïté sémantique. En effet, dans la figure 4 le terme t1 est rattaché à deux concepts différents. Sim(t1, t2) est égal à la somme de sim(C1, C2) et sim(C3,C2), C3 étant loin sémantiquement de C2, il ne sera pas pris en considération et C1 se trouve enrichi seulement par le poids de C2. Le poids sémantique de C1 devient 0,3+(0,8\*04) = 3,5. Il est à noter que le poids de t3 reste inchangé du fait qu’il n’est rattaché à aucun concept. Ceci est très important car nous permet de faire une recherche par concepts ainsi que simplement par mots clés.

La similarité entre les documents est calculée par une agrégation des similarités entre les vecteurs. Les similarités entre vecteurs sont calculés par leur cosinus :

$$\text{Cos}(V1, V2) = \frac{\sum_{i=1}^m v1_i \cdot v2_i}{\sqrt{\sum_{i=1}^m v1_i \cdot v1_i} * \sqrt{\sum_{i=1}^m v2_i \cdot v2_i}} \quad (11)$$

La similarité entre les documents revient donc à un simple calcul mathématique entre les vecteurs qui le composent.

## 5 Conclusion

Pour intégrer la notion de voisinage sémantique, nous avons utilisé une ontologie de concepts auxquels sont reliés les termes des documents. Dans un premier temps nous n'avons pris en considération que les liens de spécialisation/généralisation entre les concepts. En nous basant sur la mesure de similarité entre concepts présentée par (Wu & Palmer, 1994), nous avons proposé une nouvelle mesure telle que les descendants directs d'un concept sont considérés plus similaires au concept que ses frères. Nous avons alors défini un nouveau calcul du poids des termes SemW qui tient compte de la similarité conceptuelle entre les termes du même contexte. Le calcul de la similarité sémantique lors de l’indexation allège les traitements lors de la recherche.

Une des limites de notre approche, tient au fait que nous supposons disposer d’une ontologie de concepts reliée au corpus. Rappelons que nous nous plaçons dans le cadre de l’indexation de documents structurés, pour lesquels on peut supposer qu’il existe certaines ressources sur le vocabulaire du domaine. Pour utiliser le modèle présenté dans cet article, il suffit de disposer d’une structure hiérarchique entre concepts correspondant aux liens de spécialisation/généralisation. Cependant, nous sommes conscientes que le calcul de la mesure de similarité par restriction sur le lien « is-a » n'est pas toujours bien adapté, les autres types de liens peuvent être aussi importants dans le calcul de la similarité. Nous envisageons de travailler sur la

prise en compte d'autres types de liens comme par exemple le lien de composition. De plus, dans la réalité, les taxonomies ne sont pas toujours au même niveau de granularité, des parties peuvent être plus denses que d'autres. Ces problèmes peuvent être résolus, en partie, en associant des poids aux liens. L'affectation de ces poids peut être basée sur : les types de liens présents, la profondeur du lien dans la taxonomie et la densité du concept par ses voisins immédiats.

L'évaluation de notre mesure de similarité est nécessaire pour tester son efficacité ainsi que la pertinence d'un tel calcul lors de la phase d'indexation. Trois approches existent pour tester l'efficacité des mesures de similarité (Budanitsky & Hirst, 2001): la première étudie le cadre théorique de telles mesures par leurs propriétés et les cas qu'ils traitent, etc. Une deuxième manière consiste à comparer ces mesures par rapport à un jugement humain mais il est difficile de mettre en place de telles expérimentations qui porteraient sur un ensemble assez significatif de concepts. La troisième approche compare ces mesures par rapport à leur performance dans un cadre particulier d'une application TAL. Dans (Budanitsky & Hirst, 2001) cette application consiste à détecter et corriger des mots mal orthographiés.

Nous pouvons évaluer directement la structure d'index. Il s'agit généralement de calculer le temps d'indexation, l'espace de stockage de l'index par rapport à la taille de la base documentaire. Comme nous utilisons une ontologie, sa construction et son rattachement au corpus font partie de la phase d'indexation. Le calcul du temps de construction de l'index ne permet pas de juger de la valeur de l'index.

On peut aussi évaluer la pertinence d'un index en testant son impact sur la recherche, en utilisant les mesures de pertinence classiques de rappel et précision ou l'exhaustivité et la pertinence. La difficulté de l'évaluation de notre système est d'avoir un corpus avec des balises XML «pertinentes» (en vue d'une recherche structurée) et une ontologie associée. Quand l'ontologie est créée à partir du corpus manuellement ou par des méthodes semi-automatiques, le lien entre les termes et le concept est évident. Le problème se pose quand on dispose d'un corpus de spécialité et d'une ontologie du domaine, l'appariement entre terme et concept n'est pas toujours évident.

Dans un système de RàPC, l'évaluation de la similarité entre cas doit être guidée par la tâche, c'est à dire que les caractéristiques servant à l'évaluation de la similarité doivent être pertinentes par rapport au but du raisonnement. Dans notre système d'indexation, cela pourra se traduire par une sélection du vocabulaire d'indexation et des concepts représentés dans l'ontologie. On peut aussi imaginer une pondération associée à chaque unité sémantique. Il est évident qu'une telle indexation ne sera pas aussi fine qu'une indexation qui s'appuierait sur une représentation conceptuelle de chaque cas mais elle présente l'avantage d'être beaucoup moins coûteuse à mettre en œuvre lorsqu'on dispose au départ d'un ensemble de textes décrivant les cas. Elle facilite aussi l'intégration de nouveaux cas dans la base.

## Références

- Alhulou R., Napoli A. & Nauer E. (2003) Une mesure de similarité pour raisonner sur des documents, *Actes des Journées Nationales sur les Modèles de Raisonnement*, Paris, 27-28 novembre 2003
- Bisson G. (2000) La similarité: une notion symbolique/numérique. *Apprentissage symbolique-numérique (tome 2)*. Eds Moulet, Brito. Editions CEPADUES. pp. 169-201.
- Buckley, C., Salton, Allan, G., J. & Singhal, A. (1994) Automatic query expansion using SMART: TREC 3. In *Proceedings of TREC-3*.
- Budanitsky, A. & Hirst, G. (2001) Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, Pittsburgh, PA.
- Desmontils E. & Jacquin C. (2001) Des ontologies pour indexer un site Web. *Actes des journées francophones d'Ingénierie des Connaissances (IC'2001):131-146*
- Fellbaum C. (1998) WORDNET. An Electronic Lexical Database. In *The MIT Press*.
- Halkidi M. & Nguyen B & Varlamis I. & Vazirgiannis M. (2003) Thesus: Organising Web Document Collections based on Semantics and Clustering, *Journal on Very Large Databases, Special Edition on the Semantic Web, Novembre 2003*

- Hirst G. & St Onge D. (1998) Lexical chains as representations of context for the detection and correction of malapropisms. In *Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press*.
- Jiang J. & Conrath D. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, Taiwan, 1997.
- Krovetz R. (1997) Homonymy and polysemy in Information Retrieval. In *Proceedings of ACL/EACL'97*.
- Lenz M. (1998) Textual CBR and Information Retrieval - A Comparison. In *L. Gierl, M. Lenz (Eds.): Proc. 6th German Workshop on CBR*.
- D. Lin. (1998) An information-theoretic definition of similarity. In *Proceedings of 15th International Conference On Machine Learning, 1998*.
- Lu, X. A. & Keefer, R. B. (1994) Query expansion/reduction and its impact on retrieval effectiveness. *Overview of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225*, edited by D. K. Harman, 231-240.
- Patwardham S. (2003). Incorporating Dictionary and Corpus Information in a Measure of Semantic Relatedness, *M.S. Thesis*, August.
- Rada R., Mili H., Bicknell E., & Blettner M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17--30.
- Resnik P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal.
- Salton G. (1971). The SMART Retrieval System - experiments in automatic document processing. Perntice-Hall, Inc., Englewood Cliffs, NJ.
- Wu Z. & Palmer M. (1994) Verb Semantics and Lexical Selection, *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133-138.
- Zargayouna H. (2001) Raisonnement par similarité pour l'indexation et la recherche dans des documents multimédia. dans *Rapport interne LIMSI, N° 2001-12*, Juin 2001.
- Zargayouna H. (2004). Contexte et sémantique pour une indexation de documents semi-structurés. à paraître dans *ACM Conférence en Recherche Information et Applications, CORIA'2004*.