Extension sémantique du modèle de similarité basé sur la proximité floue des termes

Zoulikha Heddadji[*,**], Nicole Vincent[*]
Séverine Kirchner[**], Georges Stamon[*]

[*]Université René Descartes
45, rue des Saints Pères 75270 Paris CEDEX06
[**]CSTB-84, avenue Jean Jaurès Champs-sur-Marne
77421 Marne-la-Vallée CEDEX2
{zoulikha.heddadji, kirchner}@cstb.fr
{nicole.vincent, Georges.Stamon}@math-info.univ-pris5.fr

Résumé. Le modèle flou de proximité repose sur l'hypothèse que plus les occurrences des termes d'une requête se trouvent proches dans un document, plus ce dernier est pertinent. Cette mesure floue est très avantageuse dans le traitement des documents à textes courts, toutefois elle ne tient pas compte de la sémantique des termes. Nous présentons dans cet article l'intégration d'une métrique conceptuelle au modèle de proximité floue des termes pour la formalisation de notre propre modèle.

1 Introduction

Dans le cadre de la modélisation des étapes du raisonnement à partir de cas pour la réalisation d'un outil logiciel qui fera office d'un tuteur d'aide pour l'évitement des circonstances de pollution domestique exprimées dans des plaintes (Z. Bellia, 2004), nous souhaitons améliorer la méthode de tri basée sur la contiguïté des termes de la requête dans le texte d'un document source. À l'évidence, il est dans l'intérêt de l'usager du système de retrouver les cas les plus pertinents parmi les plaintes déjà traitées. Généralement, lorsqu'un utilisateur formule une requête au système, il compte retrouver les documents dont la signification du contenu se rapproche le plus de sa demande. Par exemple, pour la résolution d'une nouvelle plainte comportant le terme «couverture », il sera judicieux de retrouver les anciens cas de la mémoire archive relatifs non seulement au terme « couverture » lui-même, mais aussi aux «couettes », aux « duvets », aux « édredons », etc. Les documents contenant ces termes sont sans doute pertinents pour la plainte courante, néanmoins, ils ne seront pas sélectionnés par un modèle de recherche basé uniquement sur les occurrences directes des termes. Une solution incontournable est l'utilisation d'un réseau sémantique pour gérer le vocabulaire très variés qui peut être employé dans les plaintes. Dans l'étape de l'« élaboration » des cas en RàPC nous avons opté pour un modèle semi-structuré pour la constitution de la base. L'interface usager de notre système propose une série d'indexes sous forme de questions, dont les réponses apportent de l'information pour la description du problème. Nous avons proposé de traduire ces indexes sous forme de modèles de balise dans

un document XML, et la partie renseignée par l'utilisateur représente pour nous le contenu des balises.

Après avoir présenté les outils à l'origine de notre approche, mesure de similarité conceptuelle et modèle de proximité, nous introduisons notre approche prenant en compte les deux aspects. Le développement d'un exemple montre l'intérêt de notre méthode.

2 Les outils

Dans ce chapitre, nous rappelons brièvement la notion de mesure conceptuelle pour la gestion de la sémantique ainsi que la notion de cooccurrence floue entre les termes. Pour formaliser les relations entre les termes nous les rattachons aux concepts de WordNet (C. Fellbaum, 1998).

2.1 Aspect sémantique

Zarga et Salotti (H. Zargayouna et S. Salotti, 2004) définissent une métrique conceptuelle inspirée des travaux de Wu et Palmer (Z. Wu et M. Palmer, 1994). Elles privilégient toujours les liens père-fils par rapport aux autres liens de voisinage en adaptant la mesure de Wu et Palmer qui pénalise dans certains cas les fils d'un concept par rapport à ses frères. Elles introduisent la fonction Spec pénalisant ainsi les concepts qui ne sont pas de la même lignée. Nous illustrons cela dans l'exemple développé dans la figure 1.

$$Sim_{ZS}(C_1, C_2) = \frac{2 \times prof(PPG)}{prof(C_1) + prof(C_2) + spec(C_1, C_2)}$$

$$Spec(C_1, C_2) = dist(C_1, PPG) \times dist(C_2, PPG) \times prof_b(PPG)$$

Tel que prof(C_1) est le nombre d'arcs entre la racine de la hiérarchie et le concept C_1 en passant par le plus petit généralisant (PPG) du couple C_1, C_2. La valeur de $prof_b$(PPG) correspond au nombre maximum d'arcs qui séparent le PPG du concept Bottom.

2.2 Mesure de proximité

Cette mesure entre termes doit être mise en contexte lorsque l'on traite de documents. Le modèle vectoriel introduit par Salton (G. Salton et C. Buckley, 1998) exclut toute notion de position et de distance entre les mots. De surcroit, le modèle de Salton est mieux adapté à la codification des textes longs qu'à la codification des textes courts (A. Singhal, 1996). Compte tenu de la nature hétérogène des textes en notre possession, il est primordial d'élargir notre réflexion aux modèles de représentation adaptés à la nature de notre ressource (hétérogène). Nous avons étudié à cet effet le modèle de recherche basé sur la proximité des termes et inspiré du modèle booléen classique. L'approche de Mercier (A. Mercier et M. Beigbeder, 2005) repose sur l'hypothèse que plus les occurrences des termes d'une requête se trouvent proches dans un document de la base plus ce document est pertinent par rapport à cette requête.

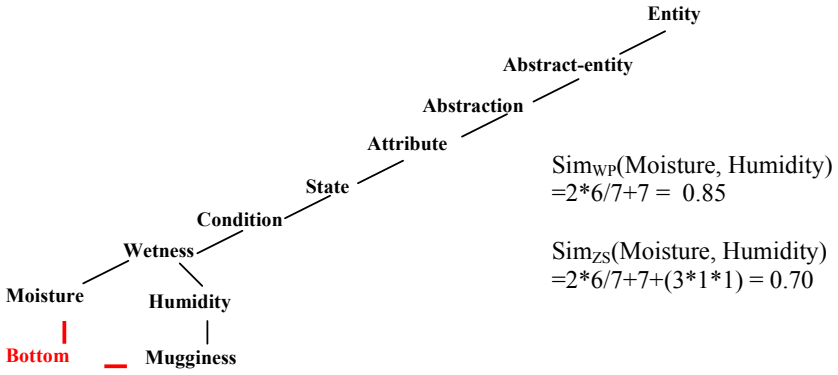

FIG.1 - Comparaison des deux modèles de similarité conceptuelle sur un exemple extrait de WordNet .

2.2.1 Proximité et pertinence locale d'un terme

La contiguïté entre termes peut être mesurée par une relation floue. Dans le cadre binaire l'opérateur $NEAR_\theta$ est évalué à 1 si A est à au moins θ mots de B. L'approche de Mercier développe la notion de proximité en «fuzzifiant» NEAR. Un document d est représenté par une suite finie de longueur l de termes t caractérisés par leur positions i. Par conséquent, la notation $d^{-1}(t)$ correspond à l'ensemble des positions prises par le terme t dans le document d.

$$\mu_{NEAR(A,B)}(d) = Max_{\substack{i\in d^{-1}(A)\\ j\in d^{-1}(B)}}\left(Max\left(\frac{k-|j-i|}{k},0\right)\right)$$

Le paramètre k est une constante[1] qui caractérise le degré d'influence d'une occurrence. Après avoir comparer deux termes et avant de comparer une requête et un document source il est nécessaire de comparer un terme et un document. Pour ce faire, la fonction μ_t^d calcule un degré de pertinence pour chaque terme t de r dans l'ensemble des positions possibles x dans d.

$$\mu_t^d(x) = Max_{i\in d^{-1}(t)}\left(Max\left(\frac{(k-|x-i|)}{k},0\right)\right)$$

2.2.2 Pertinence d'une requête par rapport à un document

Le modèle de cooccurrence directe repose sur l'hypothèse que les requêtes obéissent au modèle booléen classique, c'est à dire qu'une requête est une série de conjonctions et de disjonctions de termes. Par conséquent la pertinence locale de la requête r suit le même schéma logique entre les pertinences respectives des termes de la requête. Les opérateurs

1 Une valeur faible (5) évalue la proximité dans le cadre d'une expression, une valeur de k égale à 100 traduit la proximité dans un contexte paragraphe et ainsi de suite.

logiques sont les opérateurs classiques (AND, OR). Par exemple pour r = A AND B, la pertinence locale de r correspond à : $\mu^d_{A\,AND\,B}(x) = Min(\mu^d_A(x), \mu^d_B(x))$

Nous généralisons de manière naturelle la pertinence au niveau document en agrégeant les résultats obtenus dans l'ensemble des positions possibles.

$$Score(r,d) = \sum_{x\in[0,l-1]} \mu^d_r(x)$$

Ainsi, la similarité est obtenue en normalisant l'ensemble des scores par la longueur du document.

$$Sim(r,d) = \frac{\sum_{x\in[0,l-1]} \mu^d_r(x)}{l}$$

3 Pertinence sémantique locale

Nous apportons une extension au modèle existant en le combinant avec la mesure de similarité conceptuelle de Zarga et Salotti de la manière suivante:

$$\mu s^d_t(x) = Max_{i\in d^{-1}(Syno(t))}\left(Max\left(\frac{(k-|x-i|)\ Sim(t_i,t)}{k},0\right)\right)$$

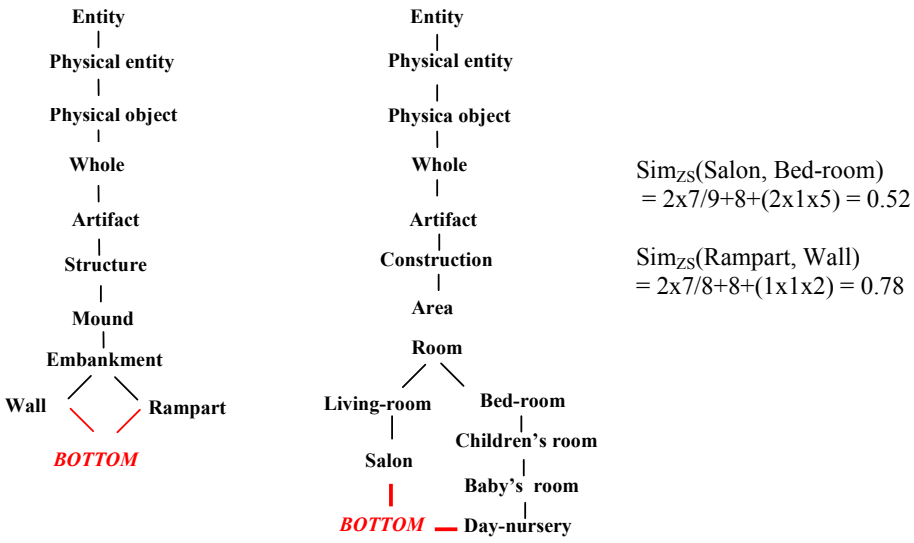

FIG. 2 – Représentation extraite du mode d'affichage récursif des concepts dans WordNet.

Par Syno(t) nous indiquons l'ensemble des termes proches sémantiquement de t. Un seuil de similarité est nécessaire pour caractériser l'ensemble de ses éléments. Nous fixons un seuil de similarité pour la valeur de $Sim(t_i,t)$ qui correspond au degré de similarité entre t et le concept auquel est rattachée la balise où il apparaît. Dans l'exemple suivant, nous comparons les deux approches. Les scores de pertinence sont calculés aux différentes positions pouvant être prises dans un intervalle d'occurrence précis de taille fixe k=10 (par exemple). Prenons en tant que exemple, la forme filtrée lemmatisée de la balise <state> (balise décrivant l'état du logement) d'une requête combinant les mots-clés suivants: r={humidity, rampart, salon}. Nous pouvons imaginer un passage de la plainte initiale exprimée de la manière suivante: «There is humidity on the rampart in the salon». Supposons qu'il existe un dossier stocké en mémoire dont la partie problème contient cet extrait:«Many moistures gleamed all on the wall of my bed-room». d={moisture, wall, bed-room}. Les termes de la requête courante, a priori, n'appartiennent pas au texte du document source, néanmoins le sens de ces deux passages est résolument le même. Le Tableau 1 montre un exemple d'application du modèle de cooccurrence directe et les résultats de notre modèle. Le symbole $ sert à indiquer les positions prises dans le texte par des termes éliminés. Nous voyons à travers cette illustration, les scores importants enregistrés par les termes de la requête aux différentes positions du document source. Par contre, leur degré de pertinence est nul en utilisant le modèle de contiguïté directe. Les valeurs de similarité conceptuelle sont calculées à l'aide du modèle de Zarga et Salotti et de WordNet (figure 2).

x	1	2	3	4	5	6	7	8	9	10
document	$	$	moisture	$	$	$	wall	bedroom	$	$
Calcul de la pertinence locale sémantique (notre approche)										
$\mu s_{humidity}^{d}(x)$	0.56	0.63	***0.70***	0.63	0.56	0.49	0.42	0.35	0.28	0.21
$\mu s_{rampart}^{d}(x)$	0.31	0.39	0.47	0.54	0.62	0.70	***0.78***	0.7	0.62	0.54
$\mu s_{salon}^{d}(x)$	0.17	0.22	0.27	0.32	0.37	0.42	0.47	***0.52***	0.47	0.42
$\mu s_{(h\ AND\ r)}^{d}(x)$	0.31	0.39	0.47	0.54	0.56	0.49	0.42	0.35	0.28	0.21
$\mu s_{(hANDr)ORs}^{d}(x)$	0.31	0.39	0.47	0.54	0.56	0.49	0.47	0.52	0.47	0.42

TAB. 1 – *Tableau comparatif des scores de pertinence locale.*

Pour l'application de cet exemple, nous nous sommes assurés que les similarités autorisées pour l'augmentation de la pertinence locale soient supérieures au degré de similarité entre le terme appartenant au document source et le terme associé à la balise où apparaît l'extrait du document source. Dans notre cas il s'agit d'un extrait de la balise <State>. Sim_{ZS}(moisture,state)=0.62<0.70, Sim_{ZS}(wall,state)=0<0.78 et Sim_{ZS}(bed-room,state)=0<0.52. La pertinence sémantique de la requête par rapport au document source dans l'exemple correspond à Sim(r,d)=0.46, alors que ce score est nul si on applique la méthode directe. Ces résultats montrent que l'extension que nous proposons augmente de manière significative la qualité des résultats. Ceci tend à prouver que l'usage des ressources sémantiques est très utile dans la phase de recherche que nous souhaitons fine.

4 Conclusion

Il est vrai que nous avons déroulé notre méthode sur un exemple, ce qui n'est pas suffisant pour savoir si nos résultats sont significatifs. L'évaluation de notre modèle est nécessaire, néanmoins la difficulté est d'avoir un corpus consistant et une ressource sémantique associée. Nous disposons actuellement d'un important corpus de plaintes résolues issues du SAMI de Liège (Système d'Analyse des Milieux Intérieurs) et du Laboratoire d'hygiène de la Ville de Paris (LHVP). Il s'est avéré indispensable d'introduire l'aspect sémantique permettant ainsi de traiter les textes réels et assez courts mis en forme.

Références

Bellia, Z.(2004). «Modélisation d'un système informatique pour la gestion des demandes d'intervention dans le domaine des ambiances intérieures: Une approche basée sur le Raisonnement à Partir de Cas». Mémoire de stage DEA-IMTC, ParisVI.

Fellbaum, C (1998). «WORDNET: An Electronic Lexical Database ». In The MIT Press.

Mercier, A et M. Beigbeder (2005).«Application de la logique floue à un modèle de recherche d'information basé sur la proximité». Dans les Actes LFA 2004, 231–237.

Salton, G. et C. Buckley (1988). «Term-Weighting Approaches in Automatic Text Retrieval». Journal of Information Processing &Management, 513–523.

Singhal, A (1996). «Pivoted length normalization». In Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR'96), 21–29.

Wu, Z. et M. Palmer (1994). «Verb Semantics and Lexical Selection ». In Proceedings of the 32nd Annual Meetings of the ACL, 133–138.

Zargayouna, H et S. Salotti (2004). «Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML». Dans Actes de la conférence IC'2004.

Summary

We present here a knowledge-based method witch combines terms proximity and the advantages of the semantic adapted to a case base annotated in XML. The information retrieval model using the fuzzy contiguity is established on the assumption that more query terms are brought closer in a document more that last is relevant. This measure is interesting; nevertheless, it does not take account of the semantic associated. The combination of a conceptual similarity measure to the proximity model seemed to us a good intuition to define a tighter model by outperforming the fuzzy proximity ranking function.