PROXIGÉNÉA : Une mesure de similarité conceptuelle

Damien DUDOGNON, Gilles HUBERT, Bachelin RALALASON
Damien.Dudognon@irit.fr, Gilles.Hubert@irit.fr, Bachelin.Ralalason@irit.fr

Institut de Recherche en Informatique de Toulouse, équipe SIG/EVI, 118 route de Narbonne, F-31062 Toulouse Cedex 9, France

Mots clefs :
Web sémantique, Recherche d'information, Ontologie, Mesure de similarité conceptuelle
Keywords:
Semantic web, Information retrieval, Ontology, Conceptual similarity measure
Palabras clave :
Escudriñar científico y tecnológico, administración del conocimiento, ingeniería del conocimiento, innovación, formalización del conocimiento, reunir de información

Résumé

La recherche d'information sémantique vise à mieux satisfaire les besoins en information des utilisateurs que ce soit sur le web ou dans les mémoires d'entreprises, en s'intéressant au sens des termes utilisés. La recherche sémantique consiste à comparer d'un point de vue sémantique (plutôt que morphologique ou syntaxique) les requêtes des utilisateurs avec les documents d'un corpus de domaine. Dans ce cadre, lorsque les concepts utilisés sont issus d'une ontologie qui définit les connaissances d'un domaine, la mise en correspondance des requêtes avec les documents est réalisée par la comparaison des concepts représentant les requêtes avec ceux qui annotent chaque document du corpus.
Dans cet article, nous proposons trois mesures de similarité sémantique pour la recherche d'information sémantique. La proposition principale est une mesure de similarité sémantique entre concepts d'une ontologie nommée ProxiGénéa basée sur le principe de proximité entre deux membres de la famille d'un arbre généalogique donné. Plus deux membres d'une famille ont d'ancêtres communs et sont proches du dernier ancêtre commun, plus ils sont proches l'un de l'autre. Ainsi, deux membres qui appartiennent à deux sous-arbres généalogiques différents sont très distants. Cette mesure a été évaluée dans le cadre d'un projet impliquant des partenaires industriels. Nous montrons qu'elle permet d'améliorer en moyenne la précision de l'ordre de 5,5% par rapport à la mesure conceptuelle proposée par Wu et Palmer (1994).

ABSTRACT

Semantic information retrieval is one of the research area which aims at better satisfying the users information needs whether it is on the Web or in the enterprise reports. The semantic search consists in comparing semantically the users' requests with the domain corpus documents. For this purpose, one of the most used techniques is the comparison of concepts representing the users' queries in relation to those annotating documents.
We present in this paper three semantic similarity measures. The main contribution is a measure between concepts called ProxiGénéa (for Genealogic Proximity) which is a measure based on the principle of nearness between two family members of a given family tree. As its name suggests, the more two members of a family have several common ancestors and the more they are closer to the last common ancestor, the more they are close the one to the other. This measure has been evaluated in the context of a project involving industrial partners. The results show a 5.5% improvement regarding precision compared to the well-known measure proposed by Wu and Palmer (1994).
Keywords: Semantic web, Information retrieval, Ontology, Conceptual similarity measure.

1. Introduction

Dans le but d'améliorer l'efficacité des systèmes de recherche d'information (SRI), une voie de recherche est la recherche d'information (RI) sémantique. La RI sémantique vise à mieux satisfaire les besoins en information des utilisateurs en exploitant le sens des termes utilisés. La recherche sémantique consiste à comparer d'un point de vue sémantique (plutôt que morphosyntaxique) les requêtes des utilisateurs avec les documents d'un corpus de domaine. La RI sémantique repose pour cela sur un processus d'indexation destiné à obtenir une représentation sémantique des documents et des requêtes. L'indexation conceptuelle (Mihalcea et al., 2000) utilise une ontologie de référence afin d'identifier les concepts permettant de décrire les documents et les requêtes. Documents et requêtes sont donc représentés par un ensemble de concepts issus d'une ontologie de référence.

Pour répondre à une requête soumise à l'utilisateur dans le cadre de la RI sémantique, la similarité entre les documents et la requête doit être évaluée de manière sémantique en se basant sur leurs annotations liées à l'ontologie. Il est donc nécessaire de disposer d'une mesure de similarité entre annotations sémantiques. Les annotations sémantiques étant basées sur les concepts de l'ontologie, une mesure de similarité entre concepts est également nécessaire.

Ainsi, les travaux présentés dans cet article consistent en la proposition des trois mesures de similarité pour la RI sémantique : une mesure de similarité entre annotations sémantiques, une mesure de similarité entre graphes de concepts (une annotation étant considérée comme un ensemble de graphes de concepts) et enfin une mesure de similarité entre concepts d'une ontologie.

L'article est organisé comme suit. La première section présente un état de l'art des principales propositions existantes en termes de similarité sémantique. La deuxième section décrit nos propositions en matière de similarité sémantique. Des résultats d'expérimentations sont présentés en section 3 et permettent une comparaison de notre mesure de similarité entre concepts à celle très utilisée proposée par Wu et Palmer (1994).

2. État de l'art

La RI sémantique rassemble différents aspects (Laublet et al., 2009) notamment la prise en compte de la sémantique dans l'appariement entre requêtes et documents. Lorsque l'indexation est réalisée en s'appuyant sur une ontologie, les documents et les requêtes sont décrits à partir des éléments de l'ontologie principalement les concepts. Le processus de RI qui mesure le degré de similarité entre une requête et un document utilise ensuite ces représentations issues de l'ontologie. Ainsi, il est nécessaire de disposer d'une mesure qui permette de calculer le degré de proximité entre deux concepts, en vue de classer les documents potentiellement pertinents par ordre décroissant de similarité.

Deux types d'approches peuvent être identifiés pour le calcul de telles distances:

- Les marges en haut, en bas, à gauche et à droite sont de 2,5cm ;
- les approches reposant uniquement sur la structure hiérarchique de l'ontologie (Rada et al., 1989) (Wu et Palmer, 1994) (Leacock et Chodorow, 1998), les approches qui incluent des informations autres que celles sur la structure hiérarchique, par exemple, des statistiques sur l'utilisation des types de concepts (Resnik, 1995) (Jiang & Conrath, 1997) (Lin, 1998).

Les approches reposant uniquement sur la structure hiérarchique reposent sur un principe de comptage d'arcs (« edge counting ») (Quillian, 1968). Rada et al. (1989) définissent la distance entre deux concepts c1 et c2 d'un réseau sémantique basé sur la relation « est un » comme le nombre minimum d'arcs à parcourir pour aller de c1 à c2.

Distedge(c1,c2)= nombre minimum d'arcs séparant c1 et c2

La mesure proposée dans (Wu et Palmer, 1994) repose, quant à elle, sur la notion de plus petit généralisant commun, c'est-à-dire le concept généralisant commun à c1 et c2 le plus éloigné de la racine. Elle est ainsi définie par :

$$Sim_{wp}(c_1, c_2) = \frac{2 \cdot depth(c)}{depth(c_1) + depth(c_2)}$$

où c est le concept qui généralise c1 et c2 et qui est le plus éloigné de la racine, et depth() est une fonction qui renvoie le nombre de nœuds entre un concept et la racine.

À partir de la mesure de distance précédente proposée par Rada et al. (1989) et de la formule de Resnik (1995), Leacock et Chodorow (1998) proposent d'évaluer la similarité entre deux concepts comme suit :

$$Sim_{lc}(c_1, c_2) = -\log\left(\frac{Dist_{edge}(c_1, c_2)}{2 \cdot D}\right)$$

où D est la profondeur maximum de la hiérarchie.

Les approches qui incluent des informations supplémentaires à celles sur structure hiérarchique cherchent le contenu informatif des nœuds. Une première proposition (Resnik, 1995) mesure, pour chaque concept, la probabilité de trouver ce concept ou un de ses descendants dans le corpus. Le contenu informatif associé à un concept c est alors défini par :

$$IC(c) = -\log(p(c))$$

où p(c) est la probabilité de trouver c ou un de ses descendants dans le corpus

avec $p(c) = \frac{freq(c)}{N}$ et $freq(c) = \sum_{t \in words(c)} count(t)$

où

- words(c) est l'ensemble des mots subsumés par le concept c
- count() renvoie le nombre d'occurrences d'un terme dans le corpus
- N est le nombre total d'occurrences des termes retrouvés dans le corpus.

Resnik (1995) définit ensuite la similarité entre deux concepts c1 et c2 par :

$$Sim_{\mathrm{Re}snik}(c_{1,}c_2) = Max(IC(c)), c \in S(c_1, c_2)$$

où S(c1,c2) est l'ensemble des concepts qui subsument c1 et c2.

Pour définir le contenu informatif d'un nœud, Seco et al. (2004) font l'hypothèse que, plus un concept a de descendants, moins il est informatif. Ils utilisent donc les hyponymes des concepts pour calculer le contenu informatif de ceux-ci :

$$IC_{wn}(c) = \frac{\log\left(\frac{hypo(c)+1}{\max_{wn}}\right)}{\log\left(\frac{1}{\max_{wn}}\right)} = 1 - \frac{\log(hypo(c))+1}{\log(\max_{wn})}$$

où hypo() est une fonction qui indique le nombre d'hyponymes d'un concept, et maxwn une constante qui indique le nombre de concepts de la taxonomie.
Des propositions combinent les deux types d'approches précédents c'est-à-dire en réutilisant les notions de contenu informatif et de plus petit ancêtre commun. Par exemple, Jiang et Conrath (1997) définissent la similarité entre deux concepts par :

$$Sim_{Jiang-Conrath}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 * IC(c)$$

Lin (1998) se base sur les propositions de Wu et Palmer (1994) et Resnik (1995) en appliquant la définition suivante :

$$Sim_{Lin}(c_1, c_2) = \frac{2 * \log P(c)}{\log P(c_1) + \log P(c_2)}$$

Les mesures de similarité basées sur les statistiques d''utilisation des concepts dans le corpus (Resnik, 1995) (Jiang & Conrath, 1997) (Lin, 1998) posent le problème du respect de la sémantique. En effet, pour ce type d'approche, la similarité entre deux concepts varie d'un corpus à l'autre, voire lorsque le corpus est modifié. Pourtant, d'un point de vue sémantique, cette similarité ne varie pas tant que l'ontologie n'est pas modifiée.

En revanche, les mesures basées sur la structure hiérarchique, comme (Wu et Palmer, 1994), prennent en compte l'aspect invariable de la sémantique des concepts dans l'ontologie. Cependant, Wu et Palmer (1994) font peu de distinction entre la relation de subsomption et la relation de fratrie. Leur mesure se base sur la proportion de concepts subsumants communs par rapport au nombre total de concepts existant dans les hiérarchies de deux concepts.

À la différence de Wu et Palmer (1994), nous faisons l'hypothèse que la similarité entre deux concepts dépend de leurs proximités respectives par rapport à l'ancêtre commun le plus spécifique. Nous proposons une nouvelle fonction de similarité entre concepts qui favorise la relation de subsomption. Cette mesure, baptisée ProxiGénéa (pour Proximité Généalogique), est détaillée dans la section suivante.

3. Similarité sémantique

Dans le cadre de nos travaux relatifs à la RI sémantique, les documents sont représentés à l'aide d'annotations, c'est-à-dire des graphes de concepts issus d'une ontologie. Une annotation peut être constituée d'un ou plusieurs graphes de concepts.

Le calcul de similarité entre annotations intervient donc à trois niveaux :

- au niveau des annotations ;
- au niveau des graphes de concepts ;
- et au niveau des concepts.

Il est donc nécessaire de disposer d'une mesure de similarité pour chacun de ces niveaux. La similarité entre annotations utilise la similarité entre graphes de concepts, cette dernière reposant sur la similarité entre concepts. Nous détaillons ces mesures dans les sections suivantes.

3.1. Similarité entre annotations

La similarité entre deux annotations est déterminée par la moyenne des similarités des graphes de concepts qui les constituent. L'algorithme de calcul est le suivant :

```
simAnnotation ← 0 ;
Pour chaque graphe de concepts G1 de l'annotation A1 faire
      simGrapheMax ← 0 ;
      calculer la similarité simGraphes G1 et G2
      Si simGraphes(G1,G2) > simGrapheMax Alors
            simGrapheMax ← simGraphe(G1,G2);
      Fin Si ;
      simAnnotation ← simAnnotation + simGrapheMax ;
Fin Pour ;
simAnnotation ← simAnnotation / NombreGraphes(A1) ;
```

Nous décrivons dans la section qui suit le principe de calcul de la similarité entre deux graphes de concepts, utilisé pour déterminer la similarité entre annotations.

3.2. Similarité de graphes de concepts

La similarité entre deux graphes est définie comme la moyenne pondérée des similarités entre les concepts qui les composent.

Deux concepts sont comparables s'ils sont descendants d'un même top-concept. Les top-concepts sont les concepts fils de la racine de l'arbre taxonomique. Il s'agit donc des concepts les plus génériques de l'ontologie.

Par ailleurs, les concepts peuvent avoir des importances différentes en fonction des applications. Ce degré d'importance est défini pour chaque top-concept de l'ontologie arbitrairement ou après une phase d'apprentissage. Le degré d'importance d'un concept correspond à celui du top-concept dont il est le descendant.

Soient :

- *G1* et *G2* deux graphes de concepts ;
- *Nœuds(G)* l'ensemble des nœuds (i.e. les concepts) du graphe *G ;*

- $G1_i$ *et* $G2_j$ des concepts appartenant respectivement aux graphes *G1 et G2 ;*
- $Coef(G_i)$ la fonction déterminant le degré d'importance d'un concept du graphe *G* ;
- et $SimConcept(G1_i,G2_j)$ la similarité entre les concepts $G1_i$ *et* $G2_j$.

La similarité entre deux graphes de concepts est définie ainsi :

$$SimGraphes(G1,G2) = \frac{\sum_{i=1}^{|Noeuds(G1)|} Coef(G1_i) \cdot \max_{j=1}^{|Noeuds(G2)|} (SimConcept(G1_i, G2_j))}{\sum_{i=1}^{|Noeuds(G1)|} Coef(G1_i)}$$

SimConcept peut être déterminée par plusieurs mesures de similarité, comme celle proposée par Wu et Palmer (1994). Dans la section suivante, nous proposons ProxiGénéa, une mesure de similarité alternative.

3.3. Similarité de concepts : ProxiGénéa

Notre mesure s'inspire du principe d'arbre généalogique familial. Nous considérons en effet que les termes d'un vocabulaire peuvent être organisés sous la forme d'un arbre hiérarchique où les termes les plus spécifiques sont rattachés aux termes plus génériques par une relation père-fils. Ainsi, la similarité entre deux concepts s'apparente à la proximité de deux membres de la famille. On parle alors de proximité généalogique.

Nous posons les deux hypothèses suivantes :

- plus deux membres quelconques de la famille ont des ancêtres en commun, plus ils sont proches l'un de l'autre ;
- l'éloignement d'un membre de la famille à partir d'un ancêtre commun influence sa distance par rapport aux autres membres de famille.

Considérons l'ontologie suivante :

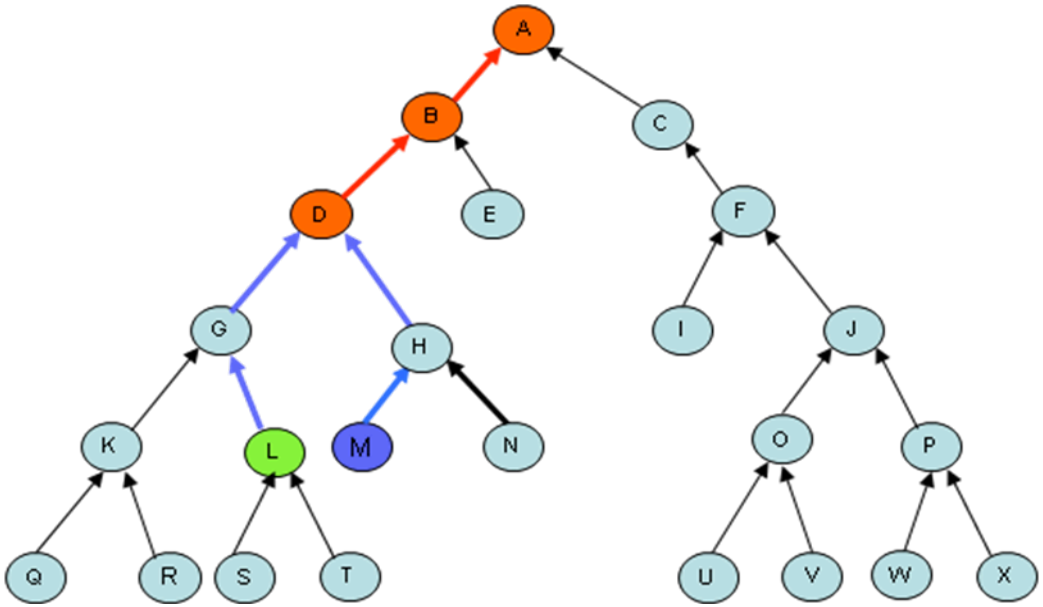

Figure 1 : Exemple d'ontologie

La similarité entre deux concepts L et M correspond au rapport du nombre d'ancêtres communs par la généalogie (i.e. l'ensemble des concepts de la racine jusqu'au concept) de ces concepts. Elle est définie de la manière suivante :

$$ProxiGénéa(L,M) = \frac{|Ancêtres(L,M)|}{|Gen(L)|} \cdot \frac{|Ancêtres(L,M)|}{|Gen(M)|} = \frac{|Ancêtres(L,M)|^2}{|Gen(L)| \cdot |Gen(M)|}$$

Avec :

- Gen(M) : l'ensemble des concepts qui entrent dans la généalogie du concept M, depuis la racine jusqu'à M.
- Ancêtres (L, M), l'ensemble des ancêtres communs des concepts L et M tels que :

$$Ancêtres(L,M) = Gen(L) \cap Gen(M)$$

L'objectif de notre mesure est de favoriser, de manière plus prononcée qu'avec la mesure de Wu et Palmer, la relation de subsomption par rapport à la relation de fratrie. A titre d'exemple, si nous considérons dans la figure 1 la relation de fratrie qui lie les concepts M et N et la relation de subsomption des concepts M et H, nous obtenons les valeurs de similarité entre concepts suivantes :

Mesure	**Sim(M,H)**	**Sim(M,N)**	**Écart fratrie/subsomption**
Wu et Palmer	0,80	0,88	9%
ProxiGénéa	0,67	0,80	16%

Tableau 1 : Comparaison de l'approche Wu et Palmer par rapport à ProxiGénéa

Ces écarts ont d'autant plus d'impact lorsqu'ils sont combinés dans une fonction de similarité plus globale, comme la fonction de similarité entre graphes de concepts.

Dans la section suivante, nous montrons au travers d'une expérimentation que ProxiGénéa est une alternative possible à la proposition de Wu et Palmer dans le domaine de la RI sémantique.

1 ÉVALUATION

Dans cette section, nous présentons l'évaluation de la fonction de similarité conceptuelle ProxiGénéa dans le cadre du projet Dynamo (DYNAMic Ontology for Information retrieval). Dans la sous section 4.1, nous détaillons le cadre expérimental Dynamo, puis dans la sous-section 4.1, nous examinons les résultats de l'expérimentation.

1.1 Cadre expérimental

L'objectif principal du projet DYNAMO est d'apporter une approche méthodologique et un ensemble d'outils logiciels permettant la construction et la maintenance de ressources ontologiques à partir de documents et l'utilisation de ces ressources pour une indexation sémantique facilitant la recherche d'information.

L'outil comporte deux modules :

- le premier est dédié à la construction et à la maintenance d'ontologies ;
- le second concerne les processus d'annotation, d'indexation et de recherche d'information.

Ces outils sont développés en tant que plugin Protégé[1] ou application autonome faisant intervenir les plugins de Protégé. Protégé a été utilisé comme plate-forme de base pour l'édition de l'ontologie de domaine et l'acquisition des connaissances.

Les mesures de similarité que nous avons définies ont été expérimentées pour la recherche sémantique de documents dans un corpus de 356 fiches de maintenance automobile.

1.2 Résultats

Nous avons comparé notre approche basée sur la similarité entre concepts ProxiGénéa à notre approche utilisant la mesure de similarité entre concepts proposée par Wu et Palmer (1994) à la place de ProxiGénéa.

Les résultats de RI sémantique sont également comparés à une recherche par mots-clés (mentionnée « classique » dans les résultats) en appliquant la fonction de similarité implémentée en standard par Lucene[2] (Gospodnetić & Hatcher 2004).

[1] http://protege.stanford.edu/

[2] Lucene a été conçu par D. Cutting (http://lucene.apache.org/).

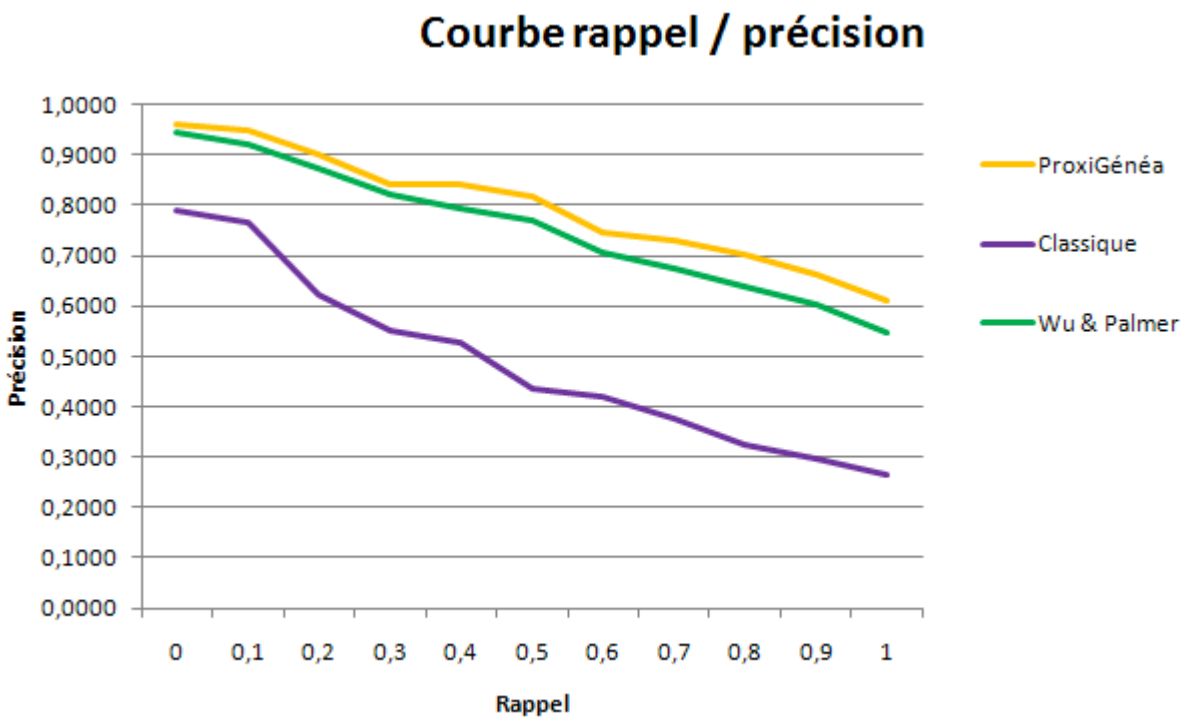

Figure 2 : Comparaison des valeurs rappel / précision

Les résultats présentés dans la figure 2 montrent la supériorité de ProxiGénéa par rapport à Wu & Palmer en matière de rappel et précision. Nous constatons aussi la nette prédominance de la RI sémantique envers la RI classique.

Le tableau 2 synthétise les précisions moyennes toutes requêtes confondues (MAP) pour les deux mesures de similarité, ProxiGénéa, Wu et Palmer, ainsi que pour la mesure classique. Ces résultats confirment l'efficacité de la mesure ProxiGénéa en termes de MAP et également des mesures sémantiques en comparaison aux approches classiques.

Mesure de similarité	MAP
ProxiGénéa	0,7718
Wu et Palmer	0,7261
Classique	0,4484

Tableau 2 : Comparaison des précisions moyenne globale (MAP)

Enfin, le tableau 3 compare les trois mesures suivant la précision obtenue à différents nombres de documents retrouvés. Ce tableau permet notamment d'évaluer la capacité des mesures à restituer des documents pertinents dans les premiers rangs de l'ensemble de résultats.

Précision	ProxiGénéa	Wu et Palmer	Classique
P1	0,9000	0,9000	0,6500
P2	0,8750	0,8250	0,5750
P5	0,7700	0,7300	0,5000
P10	0,5700	0,5550	0,3650
P15	0,4700	0,4267	0,2900
P20	0,4025	0,3700	0,2525
P30	0,3083	0,2883	0,2033
P50	0,2080	0,2050	0,1690

Tableau 3 : Valeur des précisions globales par fonction de similarité sur GenericSimilarity.

Nous pouvons constater que pour la précision à 1 document restitué (P1), les mesures Wu et Palmer et ProxiGénéa obtiennent le même résultat. Pour les autres niveaux de précision, ProxiGénéa est toujours plus efficace que Wu et palmer.

2 CONCLUSION ET PERSPECTIVES

Cet article s'inscrit dans le domaine de la recherche d'information sémantique. Dans ce cadre, documents et requêtes sont représentés par des concepts issus d'une ontologie de référence. Ainsi, les documents et les requêtes sont décrits par des annotations sémantiques, une annotation étant définie par un ensemble de graphes de concepts. Pour répondre à une requête, la recherche sémantique compare l'annotation de la requête à

celles des documents en se basant sur une mesure de similarité sémantique. Nous avons proposé dans cet article trois mesures de similarité sémantique correspondant aux trois niveaux impliqués dans les annotations : une mesure de similarité sémantique entre annotations, une mesure de similarité sémantique entre graphes de concepts et une mesure de similarité sémantique entre concepts. En fait, la mesure entre annotations repose sur la mesure entre graphes de concepts qui elle-même repose sur la mesure entre concepts.

Des expérimentations menées dans le cadre d'un projet impliquant des partenaires industriels ont montré une amélioration en moyenne de 5,5% en termes de précision par rapport à la mesure très utilisée de Wu et Palmer (1994).

Les perspectives à ces travaux sont tout d'abord de réaliser des expérimentations sur un corpus plus volumineux afin de confirmer les résultats présentés dans cet article dans un cadre plus général. Ensuite, nous devons étendre la définition de nos mesures pour prendre en compte les relations spécifiées entre les concepts dans l'ontologie.

3 REMERCIEMENTS

Les recherches présentées dans cet article s'inscrivent dans le projet ANR Dynamo (Dynamic Ontology for Information Retrieval). Les idées exprimées dans cet article sont cependant personnelles.

RÉFÉRENCES

Gospodnetić O. et Hatcher, E. « Lucene in action, *A guide to the Java search engine*. Manning, Greenwich (UK). (2004).

Jiang, J. et Conrath, D.W., « Semantic Similarity based on Corpus Statistics and Lexical Taxonomy ». *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, (1997).

Leacock, C., Miller, G. A., and Chodorow, M. 1998. Using corpus statistics and WordNet relations for sense identification. Comput. Linguist. 24, 1 (Mar. 1998), 147-165.

Laublet, P., Aussenac-Gilles, N. Camps,V., Glize, P., Hernandez, N., Maurel, H., Mbarki, M., Mothe, J., Ralalason, B., Reymonet, A., Rothenburger, B., Sellami, Z., Thomas, J., Tissaoui, A.. *Projet ANR 07 TLOG 004 01 - DYNAMO (DYNAMic Ontology for information retrieval) : État de l'art - Livrable lot 2*. Rapport de contrat, Dynamo 2.1, IRIT, décembre 2009.
Accès : ftp://ftp.irit.fr/IRIT/IC3/Rapport_lot2_integre_F7.pdf.

Lin, D., « An information-theoric definition of similarity », *In Proceedings of the 15th international conference on Machine Learning*, p. 296-304, (1998).

Mihalcea, R. et Moldovan, D. 2000. Semantic indexing using WordNet senses. Proceedings of the Acl-2000 Workshop on Recent Advances in Natural Language Processing and information Retrieval in conjunction with the 38th Annual Meeting of the Association For Computational Linguistics - Volume 11, p. 35-45, (2000).

Quillian, M.R., « Semantic Memory », M. Minsky (Ed.), Semantic Information Processing, M.I.T. Press, Cambridge, (1968).

Rada, R., Mili, H., Bicknell, E. et Blettner, M., « Development and application of a metric on semantic nets». *Systems, Man and Cybernetics*, IEEE Transactions on, 19(1): p. 17-30. (1989).

Resnik, P. « Using information content to evaluate semantic similarity in a taxonomy ». *IJCAI*, p. 448-453, (1995).

Seco, N., Veale, T. et Hayes, J. «An intrinsic information content metric for semantic similarity in Wordnet». *Proceedings of ECAI'2004*, *the 16th European Conference on Artificial Intelligence.* (2004).

Wu Z. et Palmer M., «Verb semantics and lexical selection», *Proceedings of the 23rd Annual Meetings of the Associations for Computational Linguistics*, p. 133-138, (1994).