Étude des mesures de similarité sémantique basées sur les arcs

Aly Ngoné Ngom[1]

LANI (Laboratoire d'Analyse Numérique et d'Informatique)
Université Gaston Berger
BP 234, Saint-Louis, Sénégal
alyngonengom@gmail.com

RÉSUMÉ. *Les mesures de similarité sémantique sont des fonctions très utilisées dans plusieurs domaines de l'informatique parmi lesquels nous pouvons citer le Traitement Automatique du Langage Naturel (TALN), la Bioinformatique, la Recherche d'Information... Elles permettent de déterminer la similarité entre des termes ou concepts qui n'ont aucune ressemblance syntaxique. L'objectif de ce papier est de faire une étude d'une classe particulière de mesures de similarité sémantique : les mesures basées sur les arcs. Nous ferons, dans cet article, un état de l'art de ces mesures qui aboutira à des simulations de ces dernières à l'aide d'une base de connaissances commune afin de les comparer et d'évaluer leurs performances.*

ABSTRACT. *Semantic similarity measures are functions widely used in several informatics domains among which we can mention natural language processing (NLP), Bioinformatic, Information Retrievial... They allow to determinate similarity between terms or concepts which have no syntaxic similarity. The goal of this paper is to study a particular semantic similarity group: edges based semantic similarity measures. We will do a state of art of this measures. This task will come to simulations of this measures on a common knowledge base in order to compare and to evaluate their performances.*

MOTS-CLÉS : *Similarité Sémantique, Taxonomie, Arc, Concept.*

KEYWORDS: *Semantic Similarity, Taxonomy, Arc, Concept.*

1. Encadré par :
Pr. Moussa Lo (LANI)
Dr. Fatou Kamara-Sangare (LANI)

1. Introduction

Les mesures de similarité sémantique sont des fonctions très utilisées dans plusieurs domaines de l'informatique parmi lesquels nous pouvons citer le Traitement Automatique du Langage Naturel (TALN), la Bioinformatique, la Recherche d'Information... Elles permettent de déterminer la similarité entre des termes ou concepts qui n'ont aucune ressemblance syntaxique. Leurs utilisations reposent généralement sur une bonne organisation des documents en structure hiérarchique grce à l'utilisation de bases de connaissances comme les ontologies.
Les mesures de similarité sémantique ont connu une évolution remarquable. En effet, depuis les années 90, plusieurs types de mesures ont été définis. Ces mesures peuvent être organisées en trois grandes familles que sont (Jiang et Conrath, 1997) :

– les mesures basées sur le calcul de la distance entre les concepts en prenant en compte le nombre d'arcs qui les séparent ;

– les mesures qui se basent sur la quantité d'information partagée par les concepts grce à l'utilisation de la théorie de l'information, d'o la notion de Contenu Informationnel ;

– les mesures dites hybrides qui sont basées sur la combinaison des deux familles citées plus haut ou sur l'usage de diverses techniques.

L'objectif de ce papier est de faire un état de l'art des mesures basées sur les arcs. Cette tche nous permettra de faire une étude théorique et une simulation des mesures selon des critères de comparaison.
Notre travail respectera le plan suivant : la section 2 est consacrée à une étude chronologique des mesures ; dans cette partie, neuf mesures ont été présentées ; la section 3 sera consacrée à une simulation de ces mesures sur WordNet[1]. Nous allons terminer notre étude par une conclusion et des perspectives de recherches futures.

2. Étude chronologique des mesures basées sur les arcs

Les mesures de similarité basées sur les arcs ont pour principe de compter le nombre d'arcs séparant deux concepts dans une taxonomie (Tchechmedjev, 2012). Dans cette section, nous allons nous référer à la figure 1 pour illustrer les différentes mesures étudiées. Sur cette figure, c_1 et c_2 sont deux concepts qui ont comme plus petit ancêtre commun c_3. N_1, N_2 et N_3 représentent respectivement le nombre d'arcs entre c_3 et c_1, c_3 et c_2 et c_3 et la racine. Nous allons aussi définir la profondeur d'un concept c_i dans une taxonomie comme étant le niveau de ce concept par rapport à la racine de la taxonomie. Elle est notée P_i. La profondeur totale d'une structure hiérarchique est la valeur maximale des profondeurs de l'ensemble de ces éléments. Elle est notée P_D.

1. http ://wordnet.princeton.edu

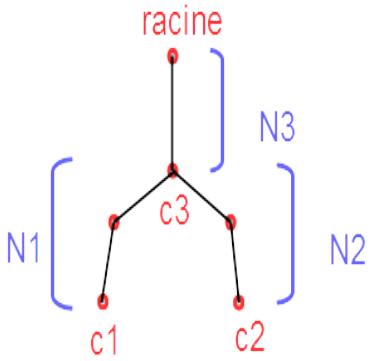

Figure 1. *Exemple de taxonomie pour les mesures de similarité basées sur les arcs*

2.1. *Mesure de Rada*

Les mesures de similarité sémantique basées sur les arcs ont été introduites par (Rada *et al.*, 1989). Elles ont été définies en fonction de la distance qui sépare deux concepts. La mesure est donnée par l'expression 1 :

$$\begin{aligned} Sim_{Rada}(c_1, c_2) &= \frac{1}{1 + dist(c_1, c_2)} \\ &= \frac{1}{1 + N_1 + N_2} \end{aligned} \quad [1]$$

2.2. *Mesure de Sussna*

(Sussna, 1993), propose une mesure de proximité sémantique qui s'appuie sur la pondération entre les concepts. La notion de proximité sémantique est plus large que la similarité sémantique (Sy, 2012) car elle ne se limite pas à la relation de subsomption (la relation de hiérarchie dans une taxonomie notée is-a).
La distance entre deux nœuds adjacents dans la taxonomie est déterminée par le calcul du poids des relations qui les relient. Elle est donnée par l'expression 2 :

$$ww(c_x, c_y) = \frac{w(c_x \rightarrow^R c_y) + w(c_y \rightarrow^{R^{-1}} c_x)}{2 \times max(p_x, p_y)}, \quad [2]$$

o $w(c_x \rightarrow^R c_y)$ et $w(c_y \rightarrow^{R^{-1}} c_x)$ représentent respectivement le poids de la relation R et le poids de la relation inverse R^{-1} existant entre les deux concepts. Le poids de la relation est donné par l'expression 3 :

$$w(c_x \rightarrow^R c_y) = max_R - \frac{max_R - min_R}{n_R(c_x)}, \quad [3]$$

avec max_R et min_R qui représentent respectivement la valeur maximale et la valeur minimale que nous pouvons associer à une relation R ; $n_R(c_x)$ représente le nombre d'arcs de $c_x \rightarrow^R c_z$.

2.3. *Mesure de Wu & Palmer*

(Wu et Palmer, 1994) ont défini une approche à base de connaissances pour la machine KBMT (*Knowledge Based Machine Translation*) dans le but de faire une traduction de l'anglais en chinois. En s'appuyant sur la figure 1, la mesure de similarité peut être exprimée par l'expression 4 :

$$\begin{aligned} Sim(c_1, c_2) &= \frac{2 \times P_3}{P_1 + P_2} \\ &= \frac{2 \times N_3}{N_1 + N_2 + 2 \times N_3} \end{aligned} \quad [4]$$

2.4. *Mesure de Leacock & Chodorow*

Leacock & Chodorow (Leacock et Chodorow, 1998) se sont inspirés des travaux de (Rada *et al.*, 1989) et de (Resnik, 1999) pour définir une mesure de similarité sémantique. Cette mesure est définie par l'expression 5 :

$$Sim_{LC}(c_1, c_2) = -\log\left(\frac{dist(c_1, c_2)}{2 \times P_D}\right) \quad [5]$$

2.5. *Mesure de Hirst & St Onge*

(Hirst et St-Onge, 1998) ont défini une mesure de proximité sémantique entre deux concepts (classes) en tenant compte des changements de direction. Deux concepts sont sémantiquement proches s'ils sont liés par un chemin qui n'est pas long (cinq arcs au maximum) et que les changements de direction ne soient pas très fréquents.
Nous parlons de changement de directions lorsque deux relations de directions différentes sont incidentes dans le chemin considéré (Hirst et St-Onge, 1998). La mesure est ainsi définie par la relation 6 :

$$\begin{aligned} Rel_{HSO}(c_1, c_2) &= C - dist(c_1, c_2) - k \times virages(c_1, c_2) \\ &= C - (N_1 + N_2) - k \times virages(c_1, c_2). \end{aligned} \quad [6]$$

Avec C et k qui sont deux constantes ($C = 8$ et $k = 1$) ; $virages(c_1, c_2)$ qui indique le nombre de changements de directions.

2.6. *Mesure de Stojanovic*

(Stojanovic *et al.*, 2001) utilise la profondeur des concepts dans une hiérarchie pour déterminer leur similarité sémantique. Cette mesure met en avant une forme généralisée de la notion de profondeur pour prendre en compte l'héritage multiple. Sa formule est donnée par l'expression 7 :

$$Sim_{Sto}(c_1, c_2) \quad = \quad \frac{P_3 + 1}{(P_1 + 1) + (P_2 + 1) - (P_3 + 1)} \qquad [7]$$

2.7. *Mesure de Zhong*

La mesure de (Zhong *et al.*, 2002) évalue l'absence de ressemblance entre les concepts. La méthode proposée est basée sur la distance entre les concepts. La mesure de similarité s'exprime ainsi :

$$Sim_{Zhong}(c_i, c_j) \quad = \quad 1 - dist(c_i, c_j). \qquad [8]$$

La distance est donnée par l'expression 9 :

$$dist(c_i, c_j) \quad = \quad \frac{1}{2^{P_{pppc}}} - \frac{1}{2^{P_i+1}} - \frac{1}{2^{P_j+1}} \qquad [9]$$

avec c_{pppc} désigne le concept communément appelé le plus petit parent commun de c_i et c_j. P_{pppc}, P_i et P_j représentent respectivent les profondeurs de c_{pppc}, c_i et c_j.

2.8. *Mesure de Zargayouna*

(Zargayouna, 2004) propose une extension de la mesure Wu & Palmer en prenant en compte le concept le plus bas de la taxonomie qu'il nomme le *bottom*. Il ajoute à la mesure de Wu & Palmer une mesure de spécificité qui prend en considération le degré de spécificité du concept. En d'autres termes, c'est le nombre d'arcs qui le séparent de *bottom*. Cette mesure s'exprime par la formule 10 :

$$\begin{cases} Sim(c_1, c_2) & = & \dfrac{2 \times P_3}{P_1 + P_2 + spec(c_1, c_2)} \\ \\ spec(c_1, c_2) & = & P_b(C) \times dist(c_1, c_3) \times dist(c_2, C) \end{cases} \qquad [10]$$

$P_b(C)$ est le nombre d'arcs qui sépare c_3 de bottom (c_3 représente le c_{pppc}).

2.9. *Mesure de Slimani*

Une mesure de similarité basée sur la mesure de Wu & Palmer est aussi proposée par (Slimani *et al.*, 2007). Cette mesure a été adoptée pour apporter des améliorations à certains problèmes de la mesure de Wu & Palmer liées à sa structure hiérarchique. La mesure proposée par (Slimani *et al.*, 2007) s'exprime par la formule 11 :

$$Sim_{tbk}(c_1, c_2) = \frac{2 \times P_3}{P_1 + P_2} \times fp(c_1, c_2) \quad [11]$$

avec

$$fp(c_1, c_2) = \begin{cases} \frac{1}{|P_1 - P_2| + 1} \text{ Si } c_1 \text{ et } c_2 \text{ sont inclus dans le même chemin ;} \\ \\ 1 \text{ sinon.} \end{cases} \quad [12]$$

2.10. *Avantages et Inconvénients*

Nous présentons, dans cette section le tableau 1 qui résume notre étude chronologique et donne les avantages et inconvients des mesures de similarité.

3. Simulation des mesures de similarité sur WordNet

Dans cette partie, nous allons faire une comparaison de certaines mesures étudiées plus haut. Il existe un certain nombre de méthodologies pour étudier et comparer des mesures de similarité. Nous pouvons citer celle qui consiste à étudier une mesure en validant ses propriétés mathématiques, celle qui consiste à inclure la mesure dans un moteur de recherche pour évaluer ses compétences ou bien la comparaison des résultats de la mesure avec le jugement humain en déterminant sa corrélation avec ce dernier. Dans nos études, nous choisissons de comparer les mesures en les comparant avec le jugement humain de (Charles et Miller, 1991). Le jugement humain de Charles & Miller consiste à faire une évaluation de la similarité entre des paires de noms en les soumettant à des sujets humains. Dans leurs expériences, (Charles et Miller, 1991) soumettent 30 paires de noms à 38 personnes pour qu'elles les affectent des notes comprises entre 0 (pas de similarité) et 4 (synonymie parfaite). La note d'une paire de nom est obtenue grce au calcul de la moyenne des 38 notes qu'elle a obtenues des sujets humains. Ces tests ont été repris par plusieurs études (ex : (Jiang et Conrath, 1997) et (Resnik, 1995)). En ce qui concerne nos études, nous nous limiterons aux 28 paires de noms utilisées dans (Jiang et Conrath, 1997). Pour effectuer cette tche, nous allons utiliser WordNet-Similarity. Notre choix a été motivé par la disponibilité de certaines mesures dans WordNet-Similarity et par son extensibilité car nous pouvons aussi y définir nos propres mesures.
Avant de procéder à la simulation, nous allons donner quelques critères des mesures de similarité sur lesquels nous nous sommes appuyés. Ces critères sont :

Mesures	Année	Avantages	Inconvénients
Rada	1989	Simple et facile à implémenter.	Ne prend pas en compte la profondeur des concepts ; ne s'adapte pas à WordNet.
Sussna	1993	Prend en compte toutes les relations de la taxonomie.	Très coteuse en terme de calcul.
Wu & Palmer	1994	Simple et facile à implémenter ; prend en compte la profondeur des concepts.	Ne donne pas une bonne similarité entre concepts voisins et concepts de la même hiérarchie.
Leacock & Chodorow	1998	Simple à implémenter.	Ne prend en compte que la relation is-a ; moins performante que Wu & Palmer sur WordNet.
Hirst & ST Onge	1998	Permet d'évaluer la similarité entre nom et verbe sur WordNet.	Limitée par des restrictions sur le nombre de chemins.
Stojanovic	2001	Prend en compte l'héritage multiple ; améliore Wu & Palmer sur WordNet.	Ne donne pas une bonne similarité entre concepts voisins et concepts de la même hiérarchie.
Zhong	2002	Donne une meilleure similarité entre "père" et "fils" qu'entre deux "frère" dans une taxonomie.	Pas de garantie de l'unicité du plus petit parent commun ($pppc$).
Zargayouna	2004	Simple et facile à implémenter ; prend en compte la profondeur des concepts et la similarité entre concepts voisins et concepts de la même hiérarchie.	Trop dépendante à l'organisation des concepts dans la Taxonomie.
Slimani	2007	Simple et facile à implémenter ; prend en compte la profondeur des concepts et la similarité entre concepts voisins et concepts de la même hiérarchie.	Trop dépendante à l'organisation des concepts dans la Taxonomie.

Tableau 1. *Récapitulation de l'étude chronologique*

– profondeur entre deux concepts (P_c) ;
– profondeur totale de la taxonomie (P_D) ;
– profondeur du plus petit parent commun des concepts ($pppc$) ;
– le nombre de virages (NV) ;
– calcul des poids des arcs ou des nœuds (CP) ;
– calcul du plus court chemin entre les concepts (PCC) ;

Cette liste de critères nous a permis de dresser le tableau 2. En nous référant au tableau 2, nous constatons que les mesures basées sur les arcs reposent généralement sur ces trois critères : P_c, PCC et $pppc$ (6 mesures utilisent P_c, 5 utilisent PCC et $pppc$) . Cette petite remarque nous a conduit à la simulation des mesures sur WordNet. Parmi les mesures qui nous intéressent, WordNet-Similarity dispose, dans sa conception,

	CRITERES DE COMPARAISONS					
Mesures	P_c	P_D	PCC	$pppc$	NV	CP
Rada	non	non	oui	non	non	non
Sussna	oui	non	oui	non	non	oui
Wu & Palmer	oui	non	non	oui	non	non
Leacock & Chodorow	non	oui	oui	non	non	non
Hirst & ST Onge	non	non	oui	non	oui	non
Stojanovic	oui	non	non	oui	non	non
Zhong	oui	non	oui	oui	non	oui
Zargayouna	oui	non	non	oui	non	non
Slimani	oui	non	non	oui	non	non

Tableau 2. *Analyse des mesures de similarité basées sur les arcs*

des mesures de Wu & Palmer, Leacock & Chodorow, Hirst & St Onge. Nous y avons ajouté les mesures de Stojanovic, Zhong et Zargayouna. La mesure de Sussna n'y a pas été ajoutée car dans sa définition, certains de ses paramètres ne sont pas justifiés. Puisque la mesure de Slimani se trouve dans le même contexte que celle de Zargayouna et Wu & Palmer, alors nous avons décider de ne pas l'implémenter pour le moment. Les résultats sont présentés dans le tableau 3. Les couples de mots qui y figurent sont extraits de WordNet. La dernière ligne du tableau nous donne les corrélations des mesures avec le jugement humain effectué par (Charles et Miller, 1991). Ainsi, nous remarquons que certaines mesures, à l'image de Zargayouna, Stojanovic et Wu & Palmer présentent les meilleurs coefficients de corrélation (respectivent 0.85, 0.80 et 0.77) par rapport à d'autres mesures telles que Leacock & Chodorow, Hirst & St Onge et Zhong (respectivement 0.76, 0.66 et 0.55. Notons aussi que Wu & Palmer et Leacock & Chodorow ont à peu près le même coefficient de corrélation.

4. Conclusion et perspectives

L'objectif de ce papier était de faire une étude d'une classe particulière de mesures de similarité sémantique : les mesures basées sur les arcs. Pour cela, nous avons, dans un premier temps, fait une étude chronologique de ces mesures, ensuite, nous avons dresser un tableau récapitulatif pour donner leurs avantages et inconvénients et enfin, une simulation de certaines mesures a été réalisée sur WordNet. Nous remarquons que les mesures basées sur les arcs présentent de bons coefficients de corrélation avec le jugement humain ; ce qui justifie leurs utilisations dans certains domaines de l'informatique. Cependant, elles sont, pour la plupart, très dépendante de la structure hiérarchique de la taxonomie. En plus, la plupart d'entre elles se limitent à la relation de subsomption. Les rares mesures qui tentent d'utiliser les autres relations définies dans les taxonomies sont confrontées à des problèmes de complexité algorithmique et sont souvent appliquées avec des restrictions. Dans nos prochaines études, nous al-

Word pair	C & M	W & P	STO	ZAR	ZHONG	LCH	HSO
car - automobile	**0,98**	1,00	1,00	1,00	1,00	0,92	1,00
gem - jewel	**0,96**	1,00	1,00	1,00	1,00	0,92	1,00
journey - voyage	**0,96**	0,95	0,92	0,95	1,00	0,75	0,25
boy - lad	**0,94**	0,93	0,89	0,93	1,00	0,75	0,31
coast - shore	**0,93**	0,92	0,87	0,92	1,00	0,75	0,25
asylum - madhouse	**0,90**	0,95	0,92	0,95	1,00	0,75	0,25
magician - wizard	**0,88**	1,00	1,00	1,00	1,00	0,92	1,00
midday - noon	**0,86**	1,00	1,00	1,00	1,00	0,92	1,00
furnace - stove	**0,78**	0,57	0,43	0,02	0,99	0,35	0,31
food - fruit	**0,77**	0,47	0,35	0,01	0,94	0,35	0,00
bird - cock	**0,76**	0,95	0,92	0,95	1,00	0,50	0,38
bird - crane	**0,74**	0,88	0,80	0,88	1,00	0,58	0,31
tool - implement	**0,74**	0,94	0,90	0,94	1,00	0,75	0,25
brother - monk	**0,71**	0,95	0,92	0,95	1,00	0,75	0,25
crane - implement	**0,42**	0,78	0,67	0,10	0,99	0,52	0,19
lad - brother	**0,42**	0,71	0,60	0,06	0,98	0,52	0,19
journey - car	**0,29**	0,20	0,16	0,00	0,75	0,20	0,00
monk - oracle	**0,28**	0,59	0,46	0,03	0,97	0,40	0,00
food - rooster	**0,22**	0,28	0,21	0,00	0,89	0,23	0,00
coast - hill	**0,22**	0,71	0,60	0,06	0,98	0,52	0,25
forest - graveyard	**0,21**	0,50	0,38	0,02	0,95	0,37	0,00
monk - slave	**0,14**	0,71	0,60	0,07	0,98	0,52	0,19
coast - forest	**0,11**	0,61	0,50	0,03	0,95	0,48	0,13
lad - wizard	**0,11**	0,71	0,60	0,06	0,98	0,52	0,19
chord - smile	**0,03**	0,44	0,33	0,01	0,94	0,32	0,00
glass - magician	**0,03**	0,53	0,41	0,02	0,97	0,40	0,00
noon - string	**0,02**	0,35	0,26	0,00	0,88	0,30	0,00
rooster - voyage	**0,02**	0,14	0,11	0,00	0,75	0,13	0,00
CORRELATION	**1,00**	0,77	0,80	0,85	0,55	0,76	0,66

Tableau 3. *Corrélation des mesures de similarité avec le jugement humain de Charles et Miller*

lons étudier la complexité algorithmique des mesures pour déterminer celles qui sont moins coteuses, ensuite, nous allons étendre notre comparaison en utilisant d'autres bases de connaissances pour déterminer la stabilité des mesures. Nous comparerons aussi ces mesures, dans nos études futures, avec les autres familles de mesures citées dans (Tchechmedjev, 2012).

5. Bibliographie

Charles, Miller G., ≪ Contextual Correlates of Semantic Similarity. ≫, *Language and Cognitive Processes, (6)*p. pp. 1 - 28, 1991.

Hirst G., St-Onge D., ≪ Lexical chains as representations of context for the detection and correction of malapropisms ≫, *In Fellbaum, 1998 The MIT Press*p. pp : 305-332, 1998.

Jiang J. J., Conrath D. W., ≪ Semantic similarity based on corpus statistics and lexical taxonomy ≫, *In Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997.

Leacock C., Chodorow M., ≪ Combining local context and WordNet sense similarity for word sense identification ≫, *In WordNet, An Electronic Lexical Database. The MIT Press*, 1998.

Rada R., Mili H., Bicknell E., Blettnerr M., ≪ Development and application of a metric on semantic nets ≫, *IEEE Transaction on systems, Man, and Cybernetics 19(1)*p. pp : 17 - 30, 1989.

Resnik P., ≪ Using Information Content to Evaluate Semantic Similarity in a Taxonomy ≫, *In Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal*, 1995.

Resnik P., ≪ Semantic Similarity in a taxonomy : An Information - Based Measure and its application to problems of Ambiguity in Natural Language ≫, *In Journal of Artificial Intelligence Reasearch 11*p. pp : 95-130, 1999.

Slimani T., Yaghlane B. B., Mellouli K., ≪ Une extension de mesure de similarit entre les concepts d'une ontologie ≫, *4th International Conference : Science of Electronic Technolgies of Information and Telecommunications, SETIT 2007 March 25 -29 2007 - Tunisia*, p. pp : 1 - 10, 2007.

Stojanovic N., Maedche A., Staab S., Stuber R., Sure Y., ≪ Seal : a framework for developing semantic portals ≫, *in Proc. of the int. conf. on Knowledge capture*, p. pp : 155-162, 2001.

Sussna M., ≪ Word sense disambiguation for free-text indexing using a massive semantic network ≫, *in Proc. of the Second International Conference on Information and Knowledge Management*, p. pp : 67-74, 1993.

Sy M. F., Utilisation d'ontologies comme support la recherche et la navigation dans une collection de document, PhD thesis, Universit Montpellier II, 2012.

Tchechmedjev A., ≪ tat de l'art : mesures de similarit smantique locales et algorithmes globaux pour la dsambigusation lexicale base de connaissances ≫, *Actes de la confrence conjointe JEP-TALN-RECITAL*, p. pp : 295-308, 2012.

Wu Z., Palmer M., ≪ Verbs semantics and lexical selection ≫, *in* U. A. f. C. L. Stroudsburg, PA (ed.), *In Proceedings of the 32nd annual meeting on ACL, volume 2 de ACL '94*, p. pp : 133-138, 1994.

Zargayouna H., ≪ Contexte et smantique pour une indexation de documents semi-structurs ≫, *LIMSI/CNRS-Universit Paris 11*p. pp : 1-15, 2004.

Zhong J., Zhu H., Li J., Yu Y., ≪ Conceptual graph matching for semantic search ≫, *in Proceedings of the 10th International Conference on Conceptual Structures (ICCS'02) (London, UK), Springer-Verlag*, p. pp : 92-106, 2002.