Sur l'évaluation de la quantité d'information d'un concept dans une taxonomie et la proposition de nouvelles mesures

Emmanuel Blanchard, Mounira Harzallah, Pascale Kuntz et Henri Briand

Laboratoire d'informatique de Nantes Atlantique
Site École polytechnique de l'université de Nantes
rue Christian Pauc
BP 50609 - 44306 Nantes Cedex 3
prénom.nom@univ-nantes.fr
http://lina.atlanstic.net/

Résumé. L'évaluation de la similarité entre concepts structurés dans une taxonomie connaît un réel essor lié au développement du web sémantique. En effet, les mesures sémantiques constituent une aide au développement et à l'exploitation des ontologies qui restent des tâches complexes dans un processus globale d'ingénierie des connaissances. Certaines mesures sémantiques sont basées sur la notion clef de contenu informationnel initialement proposé par Resnik (1995). Cet article expose notre vision du contenu informationnel à travers de nouveaux estimateurs indépendants de tout corpus. Nous généralisons la notion de contenu informationnel à un groupe de concepts et discutons la base du logarithme utilisée. Quelques propositions de mesures montrent l'analogie que l'on peut faire pour réutiliser des résultats bien connus sur une représentation ensembliste. Nous mettons en évidence la pertinence de notre approche par le biais de résultats statistiques.

1 Introduction

Le problème de l'évaluation de la similarité sémantique qui a été largement étudié en psychologie (Tversky, 1977; Rosch, 1975) connaît un nouvel essor lié au développement du web sémantique. Dans les années 70 beaucoup de recherches en classification ont été influencées par la théorie selon laquelle les classes sur un ensemble d'objets sont organisées dans une taxonomie grâce à un processus d'abstraction. La description des connaissances sur un domaine par une hiérarchie ou un graphe de concepts, appelé ontologie, est un problème central pour de nombreux systèmes développés à l'heure actuelle en ingénierie des connaissances.

Une ontologie est une spécification explicite et formelle d'une conceptualisation partagée, dont les primitives de base sont les concepts et les relations entre ces concepts (Gruber, 1993; Guarino, 1995). Associées au succès des nouveaux langages du Web sémantique (Bechhofer et al., 2004), les ontologies suscitent un intérêt croissant au sein des communautés de l'ingénierie et de la gestion des connaissances (e.g. Gruber (1993), Fürst (2004)). Cependant, malgré l'apparition d'outils d'aide à leur manipulation, leur développement et leur exploitation restent

des phases complexes dans un processus global de gestion de connaissances. En amont, une des difficultés majeures concerne la structuration des ensembles de concepts dont la taille ne cesse de croître. Et en aval, le problème consiste à rechercher efficacement des sous-ensembles de concepts à la fois en temps de calcul et en pertinence sémantique des résultats.

Pour faciliter ces tâches, le recours à des mesures sémantiques (similarités et dissimilarités) semble judicieux ; il permet de constituer une « connaissance heuristique » directement exploitable. De nombreuses mesures sémantiques ont été développées pour des objectifs applicatifs variés dans les domaines de la linguistique informatique, de l'intelligence artificielle et de la biologie aussi bien pour des objectifs académiques qu'industriel. On peut notamment faire référence à la désambigüation de mots (Resnik, 1999), à la détection et la correction de fautes d'orthographe (Budanitsky et Hirst, 2001), à la recherche d'images (Smeulders et al.), à la recherche d'information (Guarino et al., 1999)(Lee et al., 1993) ou encore à diverses applications en biologie (Lord et al., 2003)(Steichen et al., 2006).

La notion générale de similarité a été largement étudiée en ingénierie des connaissances et en apprentissage (Bisson, 2000). Si on considère un concept comme étant décrit par un sous-ensemble de caractéristiques d'importance uniforme, la similarité entre deux concepts c_i et c_j est souvent définie comme une fonction des caractéristiques communes à c_i et c_j (intersection ensembliste) et des caractéristiques qui les distinguent (différence symétrique). Le coefficient de Jaccard (1901) et celui de Dice (1945), définis pour les besoins d'études écologiques, sont probablement les plus communément utilisés.

La notion de « similarité sémantique » cherche à évaluer quantitativement la proximité sémantique entre deux concepts d'une ontologie. En terme applicatif, la structuration d'une ontologie est le fruit du consensus entre différents experts (Guarino, 1995) et, la définition d'une similarité sémantique pose des problèmes spécifiques liés intrinsèquement à l'information dont on dispose ainsi qu'à l'objectif poursuivi. Certaines mesures (e.g. (Rada et al., 1989; Wu et Palmer, 1994; Leacock et Chodorow, 1998; Blanchard et al., 2006)) exploitent uniquement la structuration des concepts (souvent taxonomique) ; d'autres, requièrent un corpus de textes sur le domaine de l'ontologie en complément (e.g. (Resnik, 1995; Jiang et Conrath, 1997; Lin, 1998)). Certains auteurs ont initié une étude théorique de ces mesures (Lin, 1998; Li et al., 2003).

De façon générale, une mesure sémantique est une application de l'ensemble $\mathcal{C} \times \mathcal{C}$ des paires de concepts d'une ontologie dans $\mathbb{R}^+$ qui permet d'évaluer quantitativement la proximité ou l'éloignement sémantique de deux concepts. Quelque soit le domaine applicatif, la pertinence de la mesure utilisée est étroitement associée à l'efficacité des algorithmes qui l'intègrent. Cependant, son choix reste un problème délicat. Pour comparer les mesures existantes, plusieurs approches complémentaires sont envisageables (Budanitsky, 1999). L'analyse formelle vise à étudier précisément leurs propriétés à la fois algorithmiques et statistiques. La comparaison avec le jugement humain analyse la corrélation entre les valeurs des mesures et les évaluations subjectives de sujets humains. L'évaluation applicative restreint l'expérimentation à un ou plusieurs cadres applicatifs bien identifiés. Dans cet article, nous menons une analyse formelle centrée sur les relations taxonomiques (généralisation/spécialisation) associées à la subsomption (is-a). La hiérarchie de subsomption qui est commune à la majorité des ontologies est généralement celle autour de laquelle s'organise une partie de la structuration des concepts (Rada et al., 1989). Des travaux récents visent à étendre les formules basées sur des hiérarchies à des graphes en prenant en compte différents types de liens simultané-

ment (Ganesan et al., 2003; Maguitman et al., 2005). Cependant, malgré leur pertinence, ces tentatives font face à des problèmes ouverts, et en pratique les approches basées sur une représentation ensembliste et celles basées sur la hiérarchie taxonomique restent les plus largement utilisées.

Dans cet article nous dressons tout d'abord un état de l'art sur les ontologies. Puis, après quelques rappels théoriques fondamentaux, nous reprenons la classification proposée par Rodriguez (2000), qui distingue trois modéles de mesures principaux pour l'évaluation de la similarité sémantique : *(1)* les modèles basés sur les caractéristiques, *(2)* ceux basés sur les relations sémantiques entre les concepts et *(3)* ceux basés sur le contenu informationnel des concepts. Le modèle que nous proposons reprend la notion de contenu informationnel initialement proposée pour exploiter un corpus de textes afin d'exploiter la structure taxonomique de l'ontologie. Nous proposons une définition générique du contenu informationnel basée sur le choix d'une distribution de probabilité ; elle consiste à extraire de la structure taxonomique des informations sur l'interprétation extensionnelle des concepts afin de rendre compte de l'importance de l'intension de chaque concept. Différentes instanciations de la définition correspondant à différentes hypothèses sur la distribution de probabilités, nous permettent de retrouver des mesures déjà publiées et d'en proposer de nouvelles. Enfin, nous complétons ce travail théorique par des résultats expérimentaux sur des échantillons de WordNet 2.0 et des jeux de tests sur lesquels des jugements humains ont été collectés.

2 État de l'art

2.1 Ontologies et taxonomies

Définitions.

En reprenant la définition de Gruber (1993), le terme d'ontologie, emprunté à la philosophie, est défini comme suit : « an ontology is a formal, explicit, specification of a shared conceptualization ». Cette définition est celle qui est très largement reprise en informatique et plus particulièrement en ingénierie des connaissances. T. Gruber définit également la conceptualisation : « A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose ». N. Guarino et P. Giaretta (1995) complètent cette définition : « An ontology is an explicit, partial account of a conceptualization ». Notons que les spécialistes des ontologies s'accordent pour considérer que les primitives cognitives de base d'une ontologie sont les concepts et les relations entre ces concepts.

La conceptualisation telle qu'elle est évoquée dans ces définitions renvoie à un modèle abstrait d'une certaine perception de la réalité, élaboré avec intension. Une ontologie est issue d'une conceptualisation partagée ou commune, c'est-à-dire qui rend compte d'un savoir consensuellement agréé par un certain collectif. Le caractère explicite d'une ontologie signifie que les primitives de base que sont les concepts et les relations sont explicitement définis. Le terme « formel » impose que l'ontologie soit explicitée dans un formalisme doté d'une sémantique formelle pour permettre sa manipulation au sein d'un système informatique. Une ontologie peut être considérée comme une spécification partielle puisqu'une conceptualisation ne peut pas toujours être entièrement formalisée dans un tel cadre, du fait d'ambiguïtés ou parce qu'aucune représentation de leur sémantique n'existe dans le langage de représentation choisi. Nous pouvons ajouter que les ontologies sont généralement développées de façon aussi

indépendante que possible du type de manipulations qui vont être opérées sur ces connaissances.

Une distinction est établie entre les ontologies de domaine portant sur des concepts renvoyant à des objets matériels ou à des concepts d'assez bas niveau (c'est-à-dire n'offrant que des possibilités limitées de raffinement) et les ontologies portant sur des concepts de haut niveau (upper ontologies) (Mizoguchi et Ikeda, 1997). Ces dernières, décrivent des notions générales comme les notions d'objet, de propriété, d'état, de valeur, de moment, d'événement, d'action, de cause et d'effet (Sowa, 2000).

M. Uschold et M. Gruninger (1996) considèrent que les ontologies varient suivant trois dimensions : le degré de formalisation de la représentation (qui varie de façon continue depuis l'informel jusqu'au rigoureusement formel), l'objectif opérationnel (communication entre utilisateurs, interopérabilité entre systèmes, application à un problème d'ingénierie comme la réutilisabilité de composants, résolution de problèmes) et le sujet (domaine de connaissance, connaissances de raisonnement, connaissances liées au modèle de représentation). En dehors du sujet qui correspond aux différents types d'ontologies évoqués précédemment, le degré de formalisation et l'objectif opérationnel modulent la définition que nous venions implicitement de dégager en relâchant sensiblement les contraintes sur le formalisme et l'indépendance aux objectifs opérationnels.

Modélisation.

Cet article se focalise sur les ontologies taxonomiques –encore appelées taxonomies– qui sont un cas particulier des ontologies dans lequel le réseau de relations est restreint à la hiérarchie de subsomption. Dans la suite, nous notons $\mathcal{O} = (\mathcal{C}, \sqsubseteq)$ une ontologie taxonomique où $\mathcal{C}$ désigne un ensemble de concepts, $\sqsubseteq$ la relation de subsomption (relation d'ordre partielle) et $c_i \sqsubseteq c_j$ la relation « c_i est plus spécifique que c_j ».

On peut faire une interprétation extensionnelle de l'ontologie en considérant les sous-ensembles d'instances attachés aux concepts :

$\mathcal{I}$, l'ensemble des instances du domaine considéré
$\mathcal{I}_{c_i} \subseteq \mathcal{I}$, l'ensemble des instances du concept c_i
$c_i \sqsubseteq c_j \Longrightarrow \mathcal{I}_{c_i} \subseteq \mathcal{I}_{c_j}$

La racine, notée c_0, d'une ontologie est souvent un concept virtuel qui est artificiellement ajouté pour regrouper sous un unique subsumant la totalité des concepts. Dans ce cas, on peut vouloir considérer que toutes les instances du domaine appartiennent à $\mathcal{I}_{c_0}$; d'où $\mathcal{I}_{c_0} = \mathcal{I}$. Cependant, dans certains cas, la racine peut être informative et $\mathcal{I}_{c_0} \subseteq \mathcal{I}$.

On peut faire une interprétation intensionnelle de l'ontologie $\mathcal{O}$ en considérant les sous-ensembles de caractéristiques (ou attributs) qui permettent de décrire les concepts :

$\mathcal{E}$, l'ensemble des caractéristiques permettant de décrire les concepts
$\mathcal{E}_{c_i} \subseteq \mathcal{E}$, l'ensemble des caractéristiques décrivant le concept c_i
$c_i \sqsubseteq c_j \Longrightarrow \mathcal{E}_{c_j} \subseteq \mathcal{E}_{c_i}$

On peut aussi avoir une vue probabiliste liée à l'interprétation extensionnelle de l'ontologie. Soit l'expérience aléatoire $\mathcal{X}$: « on prend au hasard une instance du domaine de l'ontologie considérée ». On définit $(\Omega, \mathcal{B}, P)$ un espace probabilisé associé à $\mathcal{X}$ avec Ω l'univers des possibles, $\mathcal{B}$ l'ensemble des événements et P une mesure de probabilité sur les événements de $\mathcal{B}$ tel que $\Omega = \mathcal{I}$ et $\mathcal{B} = 2^{\mathcal{I}}$. Dans cette modélisation probabiliste, $P(\mathcal{I}_{c_x})$ correspond

à la probabilité pour une instance quelconque d'appartenir à l'extension du concept c_x. Pour simplifier, $P(\mathcal{I}_{c_x})$ est désormais noté $P(c_x)$. Soulignons que par définition, $P(c_x) = \frac{|\mathcal{I}_{c_x}|}{|\mathcal{I}|}$.

2.2 Mesures sémantiques

Dans ce qui suit, nous présentons quelques résultats théoriques fondamentaux adaptés à notre formalisation puis nous proposons un état de l'art des mesures basées sur les caractéristiques des concepts, de celles basées sur les relations sémantiques et de celles utilisant le contenu informationnel.

Similarité et dissimilarité.

Rappelons brièvement les définitions classiques des notions de similarité, dissimilarité et distance (Sokal et Sneath, 1963).

Définition 1 *Une similarité sur $\mathcal{C}$ est une application $\sigma : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$ ayant les propriétés suivantes :*

$$\forall (c_i, c_j) \in \mathcal{C}^2, \quad \sigma(c_i, c_j) \geq 0 \qquad \text{(positivité)}$$
$$\forall (c_i, c_j) \in \mathcal{C}^2, \quad \sigma(c_i, c_j) = \sigma(c_j, c_i) \qquad \text{(symétrie)}$$
$$\forall (c_i, c_j, c_k) \in \mathcal{C}^3, \sigma(c_i, c_i) \geq \sigma(c_j, c_k) \qquad \text{(maximalité)}$$

Définition 2 *Une dissimilarité sur $\mathcal{C}$ est une application $\delta : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$ ayant les propriétés suivantes :*

$$\forall (c_i, c_j) \in \mathcal{C}^2, \quad \delta(c_i, c_j) \geq 0 \qquad \text{(positivité)}$$
$$\forall (c_i, c_j) \in \mathcal{C}^2, \quad \delta(c_i, c_j) = \delta(c_j, c_i) \qquad \text{(symétrie)}$$
$$\forall c_i \in \mathcal{C}, \quad \delta(c_i, c_i) = 0 \qquad \text{(indiscernabilité des identiques)}$$

Définition 3 *Une distance sur $\mathcal{C}$ est un indice de dissimilarité $\delta : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$ ayant les propriétés suivantes :*

$$\forall (c_i, c_j) \in \mathcal{C}^2, \quad \delta(c_i, c_j) = 0 \Longrightarrow c_i = c_j \qquad \text{(identité des indiscernables)}$$
$$\forall (c_i, c_j, c_k) \in \mathcal{C}^3, \quad \delta(c_i, c_j) \leq \delta(c_i, c_k) + \delta(c_k, c_j) \qquad \text{(inégalité triangulaire)}$$

Modèles basés sur les caractéristiques.

En utilisant la théorie des ensembles, Tversky (1977) a défini une mesure de similarité comme un processus d'appariement de caractéristiques. La similarité proposée est fonction des caractéristiques communes $\mathcal{E}_i \cap \mathcal{E}_j$ mais aussi de ce qui différencie les deux concepts $\mathcal{E}_i - \mathcal{E}_j$ et $\mathcal{E}_j - \mathcal{E}_i$. Le modèle de contraste de Tversky définit la similarité comme suit :

$$\sigma_{tvc}(c_i, c_j) = \Theta \cdot f(\mathcal{E}_i \cap \mathcal{E}_j) - \alpha \cdot f(\mathcal{E}_i - \mathcal{E}_j) - \beta \cdot f(\mathcal{E}_j - \mathcal{E}_i) \qquad (1)$$

où Θ, α et $\beta \geq 0$.

Les termes Θ, α et β permettent de pondérer l'importance des caractéristiques communes et différentes des deux concepts. Cela permet par ailleurs la définition de mesures asymétriques. Le modèle ratio suivant constitue la normalisation du modèle contraste précédent :

$$\sigma_{tvr}(c_i, c_j) = \frac{f(\mathcal{E}_i \cap \mathcal{E}_j)}{f(\mathcal{E}_i \cap \mathcal{E}_j) + \alpha \cdot f(\mathcal{E}_i - \mathcal{E}_j) + \beta \cdot f(\mathcal{E}_j - \mathcal{E}_i)} \tag{2}$$

où α et $\beta \geq 0$.

Nous reprenons trois similarités connues et largement utilisées que sont les similarités de Jaccard (1901), Dice (1945) et Ochiaï (1957) :

$$\sigma_{jac}(c_i, c_j) = \frac{|\mathcal{E}_i \cap \mathcal{E}_j|}{|\mathcal{E}_i| + |\mathcal{E}_j| - |\mathcal{E}_i \cap \mathcal{E}_j|} \tag{3}$$

$$\sigma_{dic}(c_i, c_j) = \frac{2 \cdot |\mathcal{E}_i \cap \mathcal{E}_j|}{|\mathcal{E}_i| + |\mathcal{E}_j|} \tag{4}$$

$$\sigma_{och}(c_i, c_j) = \frac{|\mathcal{E}_i \cap \mathcal{E}_j|}{\sqrt{|\mathcal{E}_i| \cdot |\mathcal{E}_j|}} \tag{5}$$

Une autre stratégie pour définir la similarité sémantique entre deux concepts sur la base de leur caractéristiques est de considérer leur distance euclidienne dans un espace multidimensionnel (Rips et al., 1973). Les axes de l'espace multidimensionnel représentant les caractéristiques des concepts, une distance entre deux points de l'espace rend alors compte de la distance entre les deux concepts correspondants. Krumhansl (1978) suggère de tenir compte de la densité spatiale de l'espace au niveau des deux points concernés en plus de leur distance.

Pour pouvoir utiliser ces résultats, nous devons disposer d'une caractérisation explicite des concepts de la taxonomie ainsi que d'une fonction d'évaluation de l'importance des ensembles de caractéristiques décrivant chaque concept. De telles desciptions sont souvent incomplètes voir inexistantes. Si toutefois on dispose de ces desciptions, il est difficile de définir une fonction d'évaluation qui prenne en compte l'importance relative de chaque caractéristique (leur importance pouvant être fort variable). Le cardinal ne permet pas cela et il faut par exemple avoir recours à une somme pondérée par l'importance de chaque caractéristique. Cela impose de disposer d'une masse d'information encore plus importante et difficile à produire avec fiabilité par un expert du domaine.

Modèles basés sur les relations sémantiques.

Pour construire une similarité, de nombreux travaux cherchent à exploiter la combinatoire de la structure souvent réduite à un arbre. Rada et ses collègues (1989) ont initié ces travaux en montrant que le plus court chemin (en nombre d'arcs) était un estimateur pertinent pour qualifier la proximité de deux concepts dans un réseau sémantique. Resnik (1995) expose la mesure de similarité de Leacock et Chodorow (1994; 1998) qui est définie comme l'opposé du logarithme de la mesure de Rada préalablement normalisée.

La mesure de similarité proposée par Wu et Palmer (1994) sur une taxonomie sans héritage multiple correspond à une adaptation de la mesure de Dice (1945) définie sur des ensembles. Les ensembles sont remplacés par les deux chaînes reliant c_0 à c_i et c_j. L'opérateur de cardinalité sur les ensembles est remplacé par la longueur en nombre de sommets des chemins

considérés. Ainsi, l'intersection des ensembles correspond dans cette modélisation au chemin reliant c_0 et le plus spécifique des subsumants commun à c_i et c_j. Nous pouvons ainsi proposer une définition de Wu et Palmer qui rend explicite le lien avec Dice :

$$\sigma_{wp}(c_i, c_j) = \frac{2 \cdot |SupEq(c_i) \cap SupEq(c_j)|}{|SupEq(c_i)| + |SupEq(c_j)|} \tag{6}$$

où $SupEq(c_i) = \{c | c_i \sqsubseteq c\}$ est l'ensemble contenant c_i et ses subsumants.

Dans (Resnik, 1999), le changement de notation introduit une variante de la mesure de Wu et Palmer que l'on retrouve souvent dans la littérature. En effet, la longueur d'une chaîne de c_0 à c_i (en nombre de sommets) est remplacée par la profondeur du concept c_i. Cette profondeur correspond au nombre d'arêtes et non plus au nombre de sommets. Avec ce nouvel opérateur, deux concepts n'ayant que la racine comme subsumant commun auront une similarité nulle. Il apparaît que cet opérateur est donc plus adapté lorsque l'on a une racine virtuelle du type «entity » ou «thing ». On obtient la formule suivante :

$$\sigma_{wpr}(c_i, c_j) = \frac{2 \cdot |(SupEq(c_i) - \{c_0\}) \cap (SupEq(c_j) - \{c_0\})|}{|SupEq(c_i) - \{c_0\}| + |SupEq(c_j) - \{c_0\}|} \tag{7}$$

La mesure suivante définie dans (Stojanovic et al., 2001) est l'adaptation de la mesure de Jaccard (1901) selon le schéma de Wu et Palmer. Les auteurs montrent sur un exemple qu'ils utilisent leur formule sur une taxonomie dont la racine virtuelle n'est pas représentée. Cela implique que deux concepts n'ayant que la racine comme subsumant commun ont une similarité nulle. La similarité proposée est donc la suivante :

$$\sigma_{sto}(c_i, c_j) = \frac{|(SupEq(c_i) - \{c_0\}) \cap (SupEq(c_j) - \{c_0\})|}{|(SupEq(c_i) - \{c_0\}) \cup (SupEq(c_j) - \{c_0\})|} \tag{8}$$

Des travaux récents cherchent à améliorer la précision des mesures sémantiques en considérant d'autres liens en supplément de la subsomption (Ganesan et al., 2003; Maguitman et al., 2005). D'autres propositions ont également été faites avec l'objectif de généraliser les approches à tout type de lien. Du fait de la diversité des relations qui peuvent être traitées, ces mesures peuvent avoir une sémantique qui diffère notablement des mesures précédentes. Si les mesures précédentes traitent de la ressemblance des concepts d'un point de vue sémantique, celles-ci se basent plus largement sur la ressemblance en terme de connectivité.

Sussna (1993) a défini une distance qui peut prendre en compte tous les liens dès lors que l'on définit pour chaque relation de type r un intervalle de valeurs $[min_r; max_r]$ associé au poids de cette relation. Le poids $w(c_i \rightarrow_r c_j)$ de chaque relation du type r entre deux concepts c_i et c_j est réajusté en fonction de la densité locale qui correspond au nombre de relations du type r qui partent de c_i. Pour calculer la distance entre deux concepts quelconques, il faut faire la somme des distances pour tous les liens qui composent le plus court chemin entre ces deux concepts.

La mesure proposée par Hirst et St Onge (1998) distingue quatre types de relations entre deux concepts qualifiés d'« extra-forte », « forte », « moyenne » et « faible » auxquels est associé un calcul de similarité. Par exemple, pour que deux concepts soient identifiés comme moyennement en relation, il faut qu'il existe un chemin admissible entre ceux-ci. Pour qu'un chemin soit admissible, il doit être conforme à l'un des huit modèles qu'ils ont justifié sur la base de théories psycholinguistiques.

La complexité et le manque de généricité des approches ne se limitant pas à la subsomption mettent en exergue la difficulté actuelle d'adopter de telles approches.

Modèles basés sur le contenu informationnel.

Resnik (1995) définit la similarité entre deux concepts c_i et c_j comme la quantité d'information partagée par ces deux concepts. Pour cela, la probabilité $P(c_i)$ pour une instance quelconque d'appartenir à l'extension du concept c_i est attaché à chaque concept c_i. En pratique, Resnik estime cette probabilité à partir d'un corpus de textes S par la fréquence d'occurrence de c_i dans S : $\frac{n_{c_i}}{n_{c_0}}$. Pour calculer n_{c_i}, Resnik propose de compter non seulement le nombre d'occurrences de c_i, mais aussi le nombre d'occurrences des concepts qu'il subsume. Il se base sur la définition de l'information propre de la théorie de l'information pour calculer le « contenu informationnel » d'un concept c_i :

$$IC(c_i) = -\log P(c_i) \tag{9}$$

On remarque que Resnik ne discute pas la base du logarithme, ce que nous faisons dans la suite de cet article. Il définit alors la similarité de deux concepts c_i et c_j comme la quantité d'information qu'ils partagent en proposant de prendre le « contenu informationnel » maximal des subsumants communs à c_i et c_j. Celui-ci correspond au contenu informationnel $IC(mscs(c_i, c_j))$ de leur subsumant commun le plus spécifique (formellement, $c_i \sqsubseteq mscs(c_i, c_j) \wedge c_j \sqsubseteq mscs(c_i, c_j) \wedge \nexists c_x (c_x \neq mscs(c_i, c_j) \wedge c_i \sqsubseteq c_x \wedge c_j \sqsubseteq c_x \wedge c_x \sqsubseteq mscs(c_i, c_j))$) :

$$\begin{aligned} \sigma_{res}(c_i, c_j) = & \max_{c_x \in SupEq(c_i, c_j)} IC(c_x) \\ = & IC(mscs(c_i, c_j)) \end{aligned} \tag{10}$$

avec $SupEq(c_i, c_j) = \{c | c_i \sqsubseteq c \wedge c_j \sqsubseteq c\}$

On remarque qu'avec cette mesure, un concept peut être plus similaire à lui même qu'un autre et qu'un concept peut être moins similaire à lui même que deux autres concepts différents entre eux (non respect de la maximalité). Aussi, pour deux paires de concepts ayant le même subsumant commun le plus spécifique, la mesure donne le même résultat. Ce comportement peu intuitif vient du fait que cette définition considère uniquement ce que deux concepts ont en commun mais pas ce qui les différencie. Jiang et Conrath (1997) reprennent le contenu informationnel de Resnik pour évaluer la quantité d'information de ce qui distingue les deux concepts par le biais d'une dissimilarité :

$$\delta_{jc}(c_i, c_j) = IC(c_i) + IC(c_j) - 2 \cdot IC(mscs(c_i, c_j)) \tag{11}$$

Lin (1998) propose une mesure solidement justifiée sur le plan théorique qui tient compte à la fois de l'information en commun et de celle qui les différencie avec une similarité de la forme du coefficient de Dice :

$$\sigma_{lin}(c_i, c_j) = \frac{2 \cdot IC(mscs(c_i, c_j))}{IC(c_i) + IC(c_j)} \tag{12}$$

Une nouvelle approche alternative (Seco et al., 2004; Blanchard et al., 2006) consiste à redéfinir le contenu informationnel en considérant uniquement la structure taxonomique de

l'ontologie. Cette approche évite le recours à un corpus lors du calcul du contenu informationnel ; elle se base sur l'intuition selon laquelle l'information principale extraite par l'algorithme de calcul du contenu informationnel est inhérente à la structure taxonomique et non essentiellement au corpus.

3 Un modèle basé sur la subsomption

En suivant l'approche alternative évoquée ci-dessus, nous proposons ici une définition générique du contenu informationnel basé sur le choix d'une distribution de probabilité. L'objectif est d'extraire de la structure taxonomique des informations sur l'interprétation extensionnelle des concepts permettant ainsi de rendre compte de l'importance de l'intension de chaque concept. La théorie sous-jacente fait le lien entre les divers modèles exposés dans la partie précédente.

3.1 Redéfinition du contenu informationnel

La notion de « contenu informationnel » rappelée précédemment, est commune à différentes formules proposées dans la littérature. Le calcul de $P(c_i)$ dépend de l'information dont on dispose et de l'hypothèse implicite qui est adoptée lors de la construction de l'ontologie.

Nous distinguons trois cas : *(1)* il existe un nombre statistiquement représentatif d'instances associées aux feuilles de la taxonomie $\mathcal{O}$; *(2)* il existe un corpus statistiquement représentatif ; *(3)* le calcul est uniquement basé sur la structure combinatoire de $\mathcal{O}$ et différentes hypothèses peuvent être considérées pour la distribution des instances.

Nous définissons les estimateurs $\widehat{P}(c_i)$ de $P(c_i)$ dans les différents cas en considérant dans un premier temps une taxonomie sans héritage multiple. Concernant la racine, le contenu informationnel de la racine c_0 est fixé a priori en fonction de son statut. Si la racine est virtuelle ($\mathcal{I}_{c_0} = \mathcal{I}$) alors $IC(c_0) = 0$ et par conséquent $\widehat{P}(c_0) = 1$ quelque soit la base du logarithme que nous notons b. On peut également fixer $IC(c_0)$ à une autre valeur avec dans tous les cas $\widehat{P}(c_0) = b^{-IC(c_0)}$. Nous discutons des valeurs possibles pour $IC(c_0)$ et de leur signification après avoir précisé la base du logarithme.

Cas 1 L'information disponible est $\mathcal{O}$ et $\mathcal{I}_{c_i}$ pour chaque concept feuille $c_i \in \mathcal{C}_l$ de l'ensemble des feuilles $\mathcal{C}_l \subseteq \mathcal{C}$. L'ensemble d'instances $\mathcal{I}_{c_i}$ de chaque concept non-feuille $c_i \in \mathcal{C} - \mathcal{C}_l$ peut être récursivement calculé : $\mathcal{I}_{c_i} = \bigcup_{c_x \in Fils(c_i)} \mathcal{I}_{c_x}$ avec $Fils(c_i) = \{c | c \sqsubseteq c_i \wedge \nexists c_x (c_x \neq c_i \wedge c_x \neq c \wedge c_x \sqsubseteq c_i \wedge c \sqsubseteq c_x)\}$ l'ensemble des fils de c_i. Ainsi, $\widehat{P}_1(c_i) = \frac{|\mathcal{I}_{c_i}|}{|\mathcal{I}_{c_0}|}$

Cas 2 L'information disponible est $\mathcal{O}$ et un corpus S. Alors, $\widehat{P}(c_i)$ est donné par l'estimateur de Resnik $\widehat{P}_2(c_i) = \frac{|n_{c_i}|}{|n_{c_0}|}$ pour tout $c_i \neq c_0$ avec n_{c_i} le nombre d'occurrences de c_i dans S plus le nombre d'occurrences dans S des concepts qui sont subsumés par c_i.

Cas 3 Lorsque l'information disponible est restreinte à $\mathcal{O}$, nous considérons quatre hypothèses pour la distribution des instances et, pour chacune, nous définissons un estimateur de $P(c_i)$:

3.1 Le nombre d'instances est divisé par un scalaire lors de chaque spécialisation. La probabilité pour une instance d'être associée avec un concept c_i décroît exponentiellement avec la profondeur $|SupEq(c_i) - \{c_0\}|$ de c_i :

$$\begin{aligned}\widehat{P}_{3.1}(c_0) =& b^{-IC(c_0)} \\ \widehat{P}_{3.1}(c_i) =& \frac{\widehat{P}_{3.1}(pere(c_i))}{k} \\ =& \frac{\widehat{P}_{3.1}(c_0)}{k^{|SupEq(c_i)-\{c_0\}|}}\end{aligned} \tag{13}$$

avec $pere(c_i)$ qui désigne le concept père de c_i, formellement $c_i \sqsubseteq pere(c_i) \wedge \nexists c_x (c_x \neq c_i \wedge c_x \neq pere(c_i) \wedge c_x \sqsubseteq pere(c_i) \wedge c_i \sqsubseteq c_x)$. Aussi, $k \geq 2$ est un entier fixé.

3.2 La distribution du nombre d'instances est uniforme sur l'ensemble des fils de chaque concept (Blanchard et al., 2006) :

$$\begin{aligned}\widehat{P}_{3.2}(c_0) =& b^{-IC(c_0)} \\ \widehat{P}_{3.2}(c_i) =& \frac{\widehat{P}_{3.2}(pere(c_i))}{|Fils(pere(c_i))|}\end{aligned} \tag{14}$$

3.3 La distribution du nombre d'instances est uniforme sur l'ensemble des feuilles de la taxonomie :

$$\begin{aligned}\widehat{P}_{3.3}(c_0) =& b^{-IC(c_0)} \\ \widehat{P}_{3.3}(c_i) =& \begin{cases} \frac{\widehat{P}_{3.3}(c_0)}{|\mathcal{C}_l|} & , c_i \in \mathcal{C}_l \\ \sum\limits_{c_x \in Fils(c_i)} \widehat{P}_{3.3}(c_x) & , \text{sinon} \end{cases}\end{aligned} \tag{15}$$

3.4 Les cas 3.2 et 3.3 sont complémentaires : dans le cas 3.2, l'estimateur de $P(c_i)$ dépend de la structure de $\mathcal{O}$ entre la racine c_0 et le concept c_i, tandis que dans le cas 3.3 l'estimateur dépend de la structure de $\mathcal{O}$ entre le concept c_i et ses feuilles. Pour prendre simultanément en compte les deux aspects, nous proposons un nouvel estimateur de $P(c_i)$ qui est la moyenne de $\widehat{P}_{3.2}(c_i)$ et $\widehat{P}_{3.3}(c_i)$:

$$\widehat{P}_{3.4}(c_i) = \frac{\widehat{P}_{3.2}(c_i) + \widehat{P}_{3.3}(c_i)}{2} \tag{16}$$

Dans le cas 3.4, la définition de $\widehat{P}_{3.4}$ est basée sur la moyenne arithmétique classique. Ce choix est contraint par le respect de la récursivité :

$$\widehat{P}_{3.4}(c_i) = \sum_{c_x \in Fils(c_i)} \widehat{P}_{3.4}(c_x)$$

En effet,

$$\sum_{c_x \in Fils(c_i)} \widehat{P}_{3.4}(c_x) = \frac{\widehat{P}_{3.2}(c_i) + \widehat{P}_{3.3}(c_i)}{2}$$

$$\iff \sum_{c_x \in Fils(c_i)} \frac{\widehat{P}_{3.2}(c_x) + \widehat{P}_{3.3}(c_x)}{2} = \frac{\sum_{c_x \in Fils(c_i)} \widehat{P}_{3.2}(c_x) + \sum_{c_x \in Fils(c_i)} \widehat{P}_{3.3}(c_x)}{2}$$

Extension à un groupe de concepts.

Pour l'évaluation de la similarité entre groupes de concepts (sous-ensembles de $\mathcal{C}$), nous proposons de définir une forme généralisée du contenu informationnel. Cette formule considére un ensemble de concepts comprenant les concepts dont on cherche la quantité d'information et tous les subsumants d'au moins un de ces concepts $SupEq(\mathcal{C}_i) = \{c | c_i \in \mathcal{C}_i \wedge c_i \sqsubseteq c\}$. La quantité d'information résulte du cumul des quantités d'information que chaque concept apporte vis-à-vis de ses propres subsumants. L'apport d'information d'un concept c_i correspond à sa quantité d'information $-log(P(c_i))$ à laquelle on soustrait la quantité d'information de son père $-log(pere(c_i))$. Le contenu informationnel généralisé à un groupe de concepts est le suivant :

$$\begin{aligned} IC(\emptyset) &= 0 \\ IC(\mathcal{C}_i) &= IC(c_0) + \sum_{c_x \in SupEq(\mathcal{C}_i) - \{c_0\}} -\log_b P(c_x) - (-\log_b P(pere(c_x))) \\ &= IC(c_0) + \log_b \left(\prod_{c_x \in SupEq(\mathcal{C}_i) - \{c_0\}} \frac{P(pere(c_x))}{P(c_x)} \right) \end{aligned} \quad (17)$$

Base du logarithme.

Nous précisons ici le lien qui existe entre la profondeur des concepts dans l'ontologie et le contenu informationnel. Si la profondeur est un estimateur naturel de la spécificité d'un concept, le contenu informationnel peut être considéré comme une généralisation de la profondeur permettant de tenir compte d'informations supplémentaires comme les densités locales, le nombre de feuilles subsumées ou des éléments d'information externes comme un corpus. Avec la définition proposée par Resnik, le logarithme permet au contenu informationnel de fournir des valeurs d'un ordre proportionnel aux profondeurs. Pour être doté d'une sémantique plus simple à appréhender, il est intéressant de fixer la base du logarithme de manière à ce que les valeurs obtenues soient de l'ordre de la profondeur. Il s'agit de conférer au contenu informationnel une sémantique comparable à celle de la profondeur pour qu'elle puisse être interprétée comme telle.

Définissons maintenant selon les cas précédents cette nouvelle base du logarithme. Le cas est trivial pour la première hypothèse (cas 3.1) lorsque l'on pose l'équivalence suivante : $-\log_b \frac{1}{k^{|SupEq(c_i) - \{c_0\}|}} = |SupEq(c_i) - \{c_0\}| \iff b = k$. En effet, les valeurs proportionnelles à la profondeur deviennent équivalente pour $b = k$.

Dans les autres cas, des informations supplémentaires comme les densités locales sont prises en compte dans l'évaluation du contenu informationnel. Nous cherchons à ce que globalement sur l'ontologie le contenu informationnel soit le même que dans le cas de la première hypothèse. On peut alors exprimer le contenu informationnel de l'ontologie sous la première hypothèse comme suit :

$$\begin{aligned} IC_{3.1}(\mathcal{C}_l) &= IC(c_0) + \log_k \left(\prod_{c_x \in SupEq(\mathcal{C}_l) - \{c_0\}} \frac{1}{k} \right) \\ &= IC(c_0) + |SupEq(\mathcal{C}_l) - \{c_0\}| \end{aligned}$$

Dans la seconde hypothèse (cas 3.2), les densités locales influencent le contenu informationnel et dans la troisième hypothèse (cas 3.3), c'est le nombre de feuilles des concepts qui influence le calcul tandis que dans la dernière hypothèse (cas 3.4), les deux paramètres entrent en jeu. Nous cherchons b tel que le contenu informationnel généralisé de l'ensemble des concepts de l'ontologie soit égal quelque soit l'hypothèse considérée au contenu informationnel généralisé obtenu avec la première hypothèse. On définit la base du logarithme b en posant l'équivalence des contenus informationnels :

$$\begin{aligned} & IC(\mathcal{C}_l) = IC_{3.1}(\mathcal{C}_l) \\ & \iff \log_b \left(\prod_{c_x \in SupEq(\mathcal{C}_i) - \{c_0\}} \frac{P(pere(c_x))}{P(c_x)} \right) = |SupEq(\mathcal{C}_l) - \{c_0\}| \\ & \iff b = \sqrt[|SupEq(\mathcal{C}_l) - \{c_0\}|]{\prod_{c_x \in SupEq(\mathcal{C}_i) - \{c_0\}} \frac{P(pere(c_x))}{P(c_x)}} \end{aligned}$$

On obtient finalement la moyenne géométrique des inverses des probabilités conditionnelles $P(c_x/pere(c_x))$. Cela correspond par exemple dans la seconde hypothèse à la moyenne géométrique des nombres de fils du père de chaque concept. Si la racine n'est pas virtuelle, on peut fixer par exemple $IC(c_0) = 1$ pour lui donner une unité de valeur qui revient à considérer que la racine a une profondeur de 1.

Quelques mesures remarquables.

Sur la base d'une analogie entre la description basée sur des ensembles et le contenu informationnel, on peut mettre en évidence de nouvelles mesures en utilisant celles décrites dans la section 2.2. Nous avons déjà vu que pour définir une mesure, il nous fallait définir préalablement le statut de la racine qui est soit une racine virtuelle ($IC(c_0) = 0$) soit une racine informative ($IC(c_0) \geq 1$). On doit choisir une distribution de probabilité qui définit pour chaque concept c_i un estimateur de $P(c_i)$ qui tiendra compte du statut de la racine. Il reste maintenant à définir formellement la mesure et pour cela nous proposons de nous baser sur des mesures bien connues ayant déjà été proposées sur une représentation ensembliste.

Les mesures de la section 2.2 montrent que les similarités basées sur une représentation ensembliste sont des fonctions $f(|\mathcal{E}_i|, |\mathcal{E}_j|, |\mathcal{E}_i \cap \mathcal{E}_j|)$ des cardinalités des ensembles de caractéristiques décrivant chaque concept et de leur ensemble de carcatéristiques communes. Quand

la liste des caractéristiques n'est pas disponible, une notion analogue de contenu informationnel $IC(c_i)$ évalue la quantité d'information associée à la description $\mathcal{E}_i$ d'un concept c_i. La quantité d'information associée à l'intersection $\mathcal{E}_i \cap \mathcal{E}_j$ d'une paire $\{c_i, c_j\}$ est le contenu informationnel $IC(mscs(c_i, c_j))$ de leur subsumant commun le plus spécifique (formellement, $c_i \sqsubseteq mscs(c_i, c_j) \wedge c_j \sqsubseteq mscs(c_i, c_j) \wedge \nexists c_x (c_x \neq mscs(c_i, c_j) \wedge c_i \sqsubseteq c_x \wedge c_j \sqsubseteq c_x \wedge c_x \sqsubseteq mscs(c_i, c_j))$). Nous obtenons ainsi avec le schéma des trois similarités de Jaccard, Dice et Ochiaï :

$$\sigma_{jb}(c_i, c_j) = \frac{IC(mscs(c_i, c_j))}{IC(c_i) + IC(c_j) - IC(mscs(c_i, c_j))} \tag{18}$$

$$\sigma_{db}(c_i, c_j) = \frac{2 \cdot IC(mscs(c_i, c_j))}{IC(c_i) + IC(c_j)} \tag{19}$$

$$\sigma_{ob}(c_i, c_j) = \frac{IC(mscs(c_i, c_j))}{\sqrt{IC(c_i) \cdot IC(c_j)}} \tag{20}$$

Si on utilise un corpus, la définition 19 correspond alors à la proposition de Lin (définition 12). Sous la permière hypothèse de distribution, on retrouve avec $IC(c_0) = 0$ la proposition initiale de Wu et palmer (définition 6) et avec $IC(c_0) = 1$ sa dérivée (définition 7). La définition 18 avec l'utilisation de la première hypothèse correspond à la formule de Stojanovic (définition 8). Au delà de ces quelques cas particuliers, c'est donc tout un ensemble de mesures qui est proposé par ce modèle en permettant de réutiliser les résultats connus basés sur une représentation ensembliste.

3.2 Validation du modèle

Quelque soit le domaine de recherche dans lequel une mesure est définie, se pose la question de sa validation consistant à mettre en évidence l'adéquation de sa formalisation avec la sémantique recherchée. Pour évaluer une mesure, plusieurs approches complémentaires sont envisageables (Budanitsky, 1999) :

L'analyse formelle. On vise à étudier précisément leurs propriétés théoriques (métriques, ordinales, etc.) et leur distributions statistiques.

Il s'agit d'étudier la mesure hors de tout contexte applicatif en isolant ses propriétés mathématiques et l'objet sur lequel elle porte. Cela permet de cerner l'information exploitée. En complément, on peut comparer plusieurs mesures afin de comprendre les relations qu'elles entretiennent et les corrélations qui en découlent.

La comparaison avec le jugement humain. Le principe est d'analyser la corrélation entre les valeurs de la mesure et les évaluations subjectives de sujets humains.

Les connaissances présentes dans l'ontologie ainsi que dans les autres sources d'information sont nécessairement incomplètes, imprécises et contiennent fort probablement des erreurs. On a finalement une modélisation partielle des connaissances de l'homme sur un domaine. Par conséquent, puisque les connaissances modélisées sont une vue partielle de ce qu'utilisent les individus qui formulent un jugement, la corrélation avec ces jugements humains reste un indicateur intéressant. Toutefois il est souvent délicat de tirer des conclusions d'une petite variation de corrélation.

Certaines mesures (Jiang et Conrath, 1997; Li et al., 2003) proposent la pondération de certains éléments d'information à l'aide de coefficients (α, β, etc.). Ces contributions sont définies de manière à estimer « au mieux » le jugement humain. Dans ce cadre, le jugement humain en plus d'être le référentiel de comparaison devient une source d'information de la mesure. Finalement, on se sert souvent du jeux de tests comme jeux d'apprentissage ce qui peut être abusif d'un point de vue méthodologique. Cette approche nécessite également de disposer d'un échantillon statistiquement représentatif.

La comparaison des mesures sémantiques sur le jeu de tests de Miller et Charles (1991) est très utilisée dans la littérature pour montrer l'apport des nouvelles mesures proposées. Les conclusions portent souvent sur des écarts de corrélations très faibles tandis que la corrélation linéaire de Pearson qui est utilisée est sensible au problème de représentation partielle des connaissances évoqué précédemment. Un indicateur moins ambitieux puisqu'il ne considère qu'une partie de la sémantique des mesures, mais plus fiable est la corrélation des rangs de Spearman. De plus, dans certaines applications, l'ordonnancement des paires de concepts est parfois suffisant.

L'évaluation applicative. L'expérimentation est restreinte à un cadre applicatif bien identifié.

Lors de l'évaluation d'une mesure dans l'application visée il y a parfois beaucoup de paramètres qui entrent en jeu et rendent d'autant plus difficile l'analyse des résultats. Il peut être intéressant de tester plusieurs mesures dans l'application visée afin de valider d'éventuelles conclusions théoriques prélables, et d'affiner la description du comportement attendu de la mesure dans le contexte applicatif.

Ces trois approches sont complémentaires et participent à la connaissance des mesures et à la définition de la sémantique de la mesure nécessaire à l'application visée. Par ailleurs, deux types de sémantiques bien distinctes correspondant aux notions de dissimilarité et de similarité donnent parfois lieu à des comparaisons biaisées. Dans la suite, nous limitons notre étude à quelques similarités normalisées pour pouvoir tirer des conclusions pertinentes et utiles à la validation du modèle proposé.

Dans ce qui suit, nous utilisons WordNet 2.0 qui est un référentiel conséquent nous permettant de réaliser des calculs statistiques. WordNet est un référentiel en ligne dont le développement est inspiré par des théories actuelles en psycholinguistique. Les noms anglais, les verbes et les adjectifs y sont organisés en ensembles de synonymes. Différentes relations lient ces ensembles de synonymes. WordNet vise finalement à modéliser les connaissances lexicales d'une personne dont la langue maternelle est l'anglais. Nous n'entrons pas ici dans le débat concernant la nature ontologique de WordNet.

Nous reprenons les mesures de Resnik, Jiang-Conrath ainsi que celle de Lin que nous calculons à l'aide d'un corpus conséquent, le « British National Corpus » en utilisant la méthode de comptage de Resnik et un lissage de 1 (Pedersen et al., 2004). Nous calculons également les trois mesures équivalentes à celles de Resnik, Jiang-Conrath et Lin en utilisant le contenu informationnel basé sur l'hypothèse du cas 3.3 de manière à valider les principe sous-jacent à notre modèle. Nous notons Com-H3, Diff-H3 et Dice-H3 les équivalents respectifs de Resnik, Jiang-Conrath et Lin utilisant la troisième hypothèse de distribution pour s'abstraire du corpus.

Apport du corpus.

Nous avons tiré aléatoirement un échantillon de 1000 concepts de WordNet, et calculé les similarités pour chacune des 499500 paires possibles. Du fait du tirage aléatoire et de la grande taille de WordNet, de nombreuses paires contiennent des concepts sans lien sémantique, et en conséquence donnent lieu à des valeurs de similarité nulles. Nous calculons les corrélations en omettant ces valeurs nulles. Pour approfondir l'analyse, nous avons calculé les corrélations sur différentes sous-arborescences de WordNet associées à différents thèmes (insecte, arbre) ainsi que sur les paires de concepts de jeux de tests classiques (Rubenstein et Goodenough, 1965; Miller et Charles, 1991; Finkelstein et al., 2002). Nous avons calculé la corrélation des rangs (spearman) et la corrélation linéaire (pearson) entre chaque mesure utilisant le corpus et son équivalent qui ne l'utilise pas.

	Resnik		Jiang-Conrath		Lin	
	spearman	pearson	spearman	pearson	spearman	pearson
aléatoire	0.993	0.884	0.672	0.713	0.959	0.920
arbre	0.999	0.974	0.685	0.770	0.729	0.861
insecte	0.999	0.963	0.695	0.763	0.723	0.864
M&C (1991)	0.988	0.976	0.914	0.972	0.972	0.991
R&G (1965)	0.983	0.974	0.937	0.979	0.979	0.993
F&G (2002)	0.999	0.974	0.685	0.770	0.729	0.861

TAB. 1 – *Corrélation de mesures avec leur équivalent n'utilisant pas le corpus*

Parmi les corrélations obtenues, la plus faible valeur est de $0,672$ et la majorité des valeurs se situe au dessus de $0,9$. Avec les deux jeux de tests de Miller et Charles (1991) et Rubenstein et Goodenough (1965) qui font référence dans la littérature, les valeurs sont toutes au dessus de $0,9$ et en moyenne de $0,97$. Dans l'ensemble, la table 1 montre des corrélations très élevées qui permettent de conclure qu'une grande partie de l'information dont rend compte le contenu informationnel calculé à la manière de Resnik provient de la structure taxonomique de WordNet.

En revanche, il reste quoiqu'il arrive une petite partie de l'information nécessaire au calcul de la similarité qui est extraite du corpus. Il est néanmoins difficile d'évaluer la pertinence de l'utilisation du corpus tant cela dépend du contexte applicatif et des objectifs visés. Cependant, nous pouvons utiliser comme élément d'évaluation, par ailleurs très souvent utilisé dans la littérature, la corrélation avec le jugement humain sur des jeux de tests de référence.

Pertinence du corpus.

Nous comparons maintenant les corrélations entre les six mesures précédentes et les jugements humains sur les trois jeux de tests déjà évoqués. Les résultats sont présentés dans la table 2.

L'écart maximal observé sur l'ensemble des jeux de tests entre une mesure et son équivalent n'utilisant pas le corpus est de $0,023$ et l'écart moyen est de $0,006$. Les résultats avec ou sans corpus sont donc tout à fait comparables. Ajoutons à cela que ces écarts ne sont pas plus à l'avantage des mesures avec ou sans corpus puisque les cumuls des écarts en faveur des unes

	M&C (1991)		R&G (1965)		F&G (2002)	
	spearman	pearson	spearman	pearson	spearman	pearson
Resnik (10)	0.763	0.822	0.744	0.832	0.367	0.376
Com-H3 (10,15)	0.740	0.800	0.741	0.832	0.371	0.386
Jiang-Conrath (11)	-0.788	-0.846	-0.784	-0.855	-0.342	-0.347
Diff-H3 (11,15)	-0.802	-0.848	-0.797	-0.855	-0.341	-0.344
Lin (12)	0.754	0.833	0.761	0.853	0.363	0.372
Dice-H3 (19,15)	0.754	0.831	0.764	0.857	0.361	0.376

TAB. 2 – *Corrélations de Spearman et Pearson avec des jugements humains*

moins le cumul des écarts en faveur des autres est de $0,002$ qui divisé par les 18 écarts pris en compte donne une moyenne de $0,0001$ en faveur des mesures avec utilisation du corpus.

Si l'on restreint la comparaison aux valeurs obtenues avec le coefficient de Spearman (ou de Pearson), les écarts sont toujours aussi peu significatifs. La table 2 montre donc que sur la base des corrélations avec les jugements humains, l'information supplémentaire extraite du corpus ne permet pas d'améliorer l'évaluation de la similarité.

4 Conclusion

Le concept de similarité est fondamental dans beaucoup de domaines (e.g. classification, IA, psychologie, ...). A l'origine, les mesures sont souvent construites pour atteindre des objectifs précis. Cependant, quelques mesures (e.g. Jaccard (1901), Dice (1945)) ont montré leur pertinence dans des applications très diverses. Aujourd'hui les similarités suscitent un regain d'intérêt du fait du succès des ontologies en ingénierie des connaissances. Dans ce cadre, les mesures les plus utilisées pour quantifier les proximités entre paires de concepts sont les mesures basées sur les liens sémantiques dont le calcul intègre ou non un corpus de textes en complément.

Dans ce papier, nous avons mis en évidence les liens existants entre diverses mesures. Nous avons présenté différentes formules dépendant du « contenu informationnel » et nous avons proposé une nouvelle façon d'appréhender cette notion pour exploiter efficacement l'information structurelle renfermée par la taxonomie. De plus, nous avons généralisé le contenu informationnel à un groupe de concepts et fixé la base du logarithme pour lui donner une sémantique simple à interpréter.

Au-delà des résultats théoriques de ce papier, nous croyons qu'il contribue à une meilleure compréhension du comportement des similarités sur les ontologies. En pratique, le choix d'une similarité est critique du fait que les résultats dépendent souvent de ce choix.

Références

Bechhofer, S., F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, et L. A. Stein (2004). Owl web ontology language reference. http ://www.w3.org/TR/2004/REC-owl-ref-20040210/.

Bisson, G. (2000). *La similarité : une notion symbolique/numérique. Apprentissage symbolique-numérique (tome 2)*, Chapitre XX. Cépaduès.

Blanchard, E., P. Kuntz, M. Harzallah, et H. Briand (2006). A tree-based similarity for evaluating concept proximities in an ontology. In *Proc. of 10th conf. of the Int. Fed. of Classification Soc.*, pp. 3–11. Springer.

Budanitsky, A. (1999). Lexical semantic relatedness and its application in natural language processing. Technical report, Computer Systems Research Group - University of Toronto.

Budanitsky, A. et G. Hirst (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*.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology 26*(3), 297–302.

Finkelstein, L., E. Gabrilovich, Y. Matias, G. W. E. Rivlin, Z. Solan, et E. Ruppin (2002). Placing search in context : The concept revisited. *ACM Transactions on Information Systems 20*(1), 116–131.

Fürst, F. (2004). *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation*. Ph. D. thesis, Ecole polytechnique de l'université de Nantes.

Ganesan, P., H. Garcia-Molina, et J. Widom (2003). Exploiting hierarchical domain structure to compute similarity. *ACM Trans. on Information Systems 21*(1), 64–93.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition 5*(2), 199–220.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *Human-Computer Studies 43*(5/6), 625–640.

Guarino, N. et P. Giaretta (1995). Ontologies and knowledge bases : Towards a terminological clarification. In *Towards Very Large Knowledge Bases : Knowledge Building and Knowledge Sharing*, pp. 25–32. IOS Press.

Guarino, N., C. Masolo, et G. Vetere (1999). Ontoseek : Content-based access to the web. *IEEE Intelligent Systems 14*(3), 70–80.

Hirst, G. et D. St-Onge (1998). Lexical chains as representation of context for the detection and correction of malaproprisms. In C. Fellbaum (Ed.), *WordNet : An electronic lexical database*, pp. 305–332. MIT Press.

Jaccard, P. (1901). Distribution of the alpine flora in the dranse's basin and some neighbouring regions (in french). *Bulletin de la Soc. Vaudoise Sci. Nat.* (37), 241–272.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of int. conf. on Research in Computational Linguistics*, pp. 19–33.

Krumhansl, C. (1978). Concerning the applicability of geometric models to similarity data : The interrelationship between similarity and spatial density. *Psychological Review 85*(5), 446–463.

Leacock, C. et M. Chodorow (1994). Filling in a sparse training space for word sense identification.

Leacock, C. et M. Chodorow (1998). Combining local context and wordnet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet : An electronic lexical database*, pp. 265–283. MIT Press.

Lee, J. H., M. H. Kim, et Y. J. Lee (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation 49*(2), 188–207.

Li, Y., Z. A. Bandar, et D. McLean (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Trans. on Knowledge and data engineering 15*(4), 871–882.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the 15th int. conf. on machine learning*, pp. 296–304. Morgan Kaufmann.

Lord, P., R. Stevens, A. Brass, et C. Goble (2003). Investigating semantic similarity measures across the gene ontology : the relationship between sequence and annotation. *Bioinformatics 19*(10), 1275–1283.

Maguitman, A. G., F. Menczer, H. Roinestad, et A. Vespignani (2005). Algorithmic detection of semantic similarity. In *Proc. of the 14th int. conf. on world wide web*, pp. 107–116. ACM Press.

Miller, G. et W. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Mizoguchi, R. et M. Ikeda (1997). Towards ontology engineering. In *Proceedings of the Joint Pacific Asian Conference on Expert Systems*.

Ochiaï, A. (1957). Zoogeographic studies of the soleoid fishes found in japan and its neighbouring regions. *Bulletin of the Japanese Society for Scientific Fisheries 22*, 526–530.

Pedersen, T., S. Patwardhan, et J. Michelizzi (2004). Wordnet : :similarity - measuring the relatedness of concepts. In *In proc. of the Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 38–41.

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

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th int. Joint conf. on Artificial Intelligence*, Volume 1, pp. 448–453.

Resnik, P. (1999). Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research 11*, 95–130.

Rips, L. J., E. J. Shoben, et E. E. Smith (1973). Semantic distance and the verification of semantic relations. *Verbal Learning and Verbal Behavior 12*, 1–20.

Rodríguez, A. (2000). *Semantic Similarity Among Spatial Entity Classes*. Ph. D. thesis, Department of Spatial Infromation Science and Engineering University of Maine.

Rosch, E. (1975). Cognitive representations of semantic categories. *Experimental Psychology : Human Perception and Performance 1*, 303–322.

Rubenstein, H. et J. Goodenough (1965). Contextual correlates of synonymy. *Communications of the ACM 8*(10), 627–633.

Seco, N., T. Veale, et J. Hayes (2004). An intrinsic information content metric for semantic

similarity in wordnet. In *Proc. of the 16th european conf. on artificial intelligence*, pp. 1089–1090.

Smeulders, A., M. Worring, S. Santini, A. Gupta, et R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence 22*(12).

Sokal, R. R. et P. H. Sneath (1963). *Principles of numerical taxonomy*. W. H. Freeman and Co.

Sowa, J. F. (2000). Ontology, metadata and semiotics. In S.-V. LNCS (Ed.), *Proceedings of the 8th International Conference on Conceptual Structures*, Volume 1867, pp. 55–81.

Steichen, O., C. D.-L. Bozec, M. Thieu, E. Zapletal, et M.-C. Jaulent (2006). Computation of semantic similarity within an ontology of breast pathology to assist inter-observer consensus. *Computers in Biology and Medicine 36*(7-8), 768–788.

Stojanovic, N., A. Maedche, S. Staab, R. Studer, et Y. Sure (2001). Seal : a framework for developing semantic portals. In *Proc. of the int. conf. on Knowledge capture*, pp. 155–162.

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proc. of the Second International Conference on Information and Knowledge Management*, pp. 67–74.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352.

Uschold, M. et M. Gruninger (1996). Ontologies : Principles, methods and applications. *Knowledge Engineering Review 11*(2), 93–155.

Wu, Z. et M. Palmer (1994). Verb semantics and lexical selection. In *Proc. of the 32nd annual meeting of the associations for Comp. Linguistics*, pp. 133–138.

Summary

The evaluation of similarity between two concepts structured in a taxonomy has known a noticeable renewed interest linked to the development of the semantic Web. In fact, the developpement and the exploitation of ontologies which remains complex tasks in the global process of knowledge engineering could be simplify by the use of semantic measures. Several semantic measures are based on the key notion of information content initially proposed by Resnik (1995). This paper state our vision of the information content by the way of new estimators independent from any corpus. We generalize the information content notion to a group of concepts and we discuss about the logarithm base used. Several propositions of measures show the analogy which is made to reuse the well-known results on a set-based representation. We underline the relevance of our approach by the way of statistical results.