Utilisation des fonctions de croyance pour l'estimation du contenu informationnel des concepts d'une ontologie

Sébastien Harispe, Abdelhak Imoussaten, François Trousset, Jacky Montmain

Utilisation des fonctions de croyance pour l'estimation du contenu informationnel des concepts d'une ontologie

On the consideration of a bring-to-mind model for computing the Information Content of concepts defined into ontologies

S. Harispe[1] A. Imoussaten[1] F. Trousset[1] J. Montmain[1]

[1] Centre de Recherche LGI2P/Ecole des mines d'Alès

Parc scientifique G. Besse, 30035 Nîmes cedex 1, France, prenom.nom@mines-ales.fr

Résumé : Les ontologies sont le support de nombreuses applications basées sur l'exploitation de connaissances expertes. Elles sont utilisées en particulier pour estimer le contenu informationnel (*IC*) des concepts clés d'un domaine : une notion fondamentale dont dépendent diverses analyses basées sur les ontologies, *e.g.* les mesures sémantiques. Cet article propose de nouveaux modèles d' *IC* basés sur la théorie des fonctions de croyance. Ces modèles ont pour objet de remédier à une limitation des modèles classiques qui ne tiennent pas compte de *l'Hypothèse d'Inférence Inductive* (*HII*) pourtant intuitivement utilisée par l'homme. Dans les modèles classiques d' *IC* , les occurrences d'un concept (*e.g. Maths*) ont une influence sur l' *IC* des concepts plus généraux subsumant le concept (*e.g. Sciences*) ; en revanche, elles n'affectent en rien l' *IC* d'un concept subsumé (*e.g. Algebra*). C'est ce comportement que se propose de prendre en compte l'*HII*. Les propriétés attendues de notre modèle d' *IC* donnent les contraintes mathématiques à respecter lors de sa construction. Des évaluations empiriques viennent vérifier qu'il a également un comportement des plus satisfaisants pour les cas d'usage les plus classiques d' *IC* .

Mots-clés :
Contenu informationnel, ontologie, fonctions de croyance, similarité sémantique.

Abstract: Ontologies are core elements of numerous applications that are based on computer-processable expert knowledge. They can be used to estimate the Information Content (*IC*) of the key concepts of a domain: a central notion on which depend various ontology-driven analyses, *e.g.* semantic measures. This paper proposes new *IC* models based on the belief functions theoretical framework. These models overcome limitations of existing *ICs* that do not consider the inductive inference assumption intuitively assumed by human operators, *i.e.* that occurrences of a concept (*e.g. Maths*) not only impact the IC of more general concepts (*e.g. Sciences*), as considered by traditional *IC* models, but also the one of more specific concepts (*e.g. Algebra*). Interestingly, empirical evaluations show that, in addition to modelling the aforementioned assumption, proposed *IC* models compete with best state-of-the-art models in several evaluation settings.

Keywords:
Information content, ontology, belief functions; semantic similarity.

1 Problématique

Les ontologies sont le support de nombreuses applications basées sur de la connaissance experte, *e.g.* systèmes d'information médicale ou d'aide à la décision clinique [1]. En particulier, elles fournissent des taxonomies qui établissent un ordre partiel des concepts clefs d'un domaine (*e.g.* classification de maladies). En définissant des relations de généralisation/spécialisation (*i.e.* les relations d'hyperonymie/hyponymie) entre les concepts, ces taxonomies traduisent des consensus cognitifs sur une hiérarchie d'abstraction structurant les concepts du domaine. Elles sont donc très utiles pour concevoir des systèmes en Intelligence Artificielle et largement utilisées en Recherche d'Information (RI), Traitement Automatique du Langage Naturel (TALN) et raisonnement approché pour ne citer que quelques usages. Les taxonomies permettent d'analyser les propriétés intrinsèques et contextuelles des concepts. En effet, en analysant leurs topologies et diverses informations propres à l'usage d'un concept, plusieurs auteurs ont proposé des modèles qui

tirent avantage de ces taxonomies pour estimer *l'informativité* ou contenu informationnel (*IC*) de concepts [2]. Ces modèles sont conçus pour apprécier l'idée intuitive et conceptuelle que l'homme peut se faire de *l'informativité* d'un concept. Par exemple, la plupart des gens seront d'accord pour dire que le concept *Algèbre* est plus informatif que le concept *Mathématiques* dans le sens où savoir que Lucie étudie l'algèbre est plus informatif que de savoir qu'elle étudie les mathématiques. Dans ce cas, la conjecture semble naturelle puisque le premier fait implique le second étant donné que l'algèbre est un domaine d'étude particulier des mathématiques. Disposer d'estimateurs d' *IC* précis revêt une importance particulière dès lors que l'on cherche à étendre l'usage des ontologies à des applications qui font appel au raisonnement approché. En effet, nombre d'analyses basées sur des ontologies, *e.g*. calculer des similarités entre concepts, dépendent fortement de la qualité de ces estimateurs. Ainsi, les modèles d' *IC* jouent-ils un rôle fondamental dans la définition de mesures sémantiques, largement utilisées en RI, TALN et inférence de connaissances.

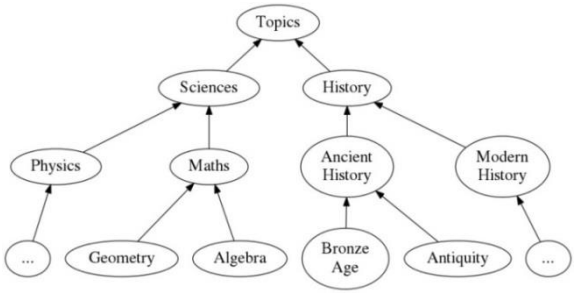

Figure 1 - Taxonomie de sujets

Une ontologie peut par exemple servir à caractériser les centres d'intérêt d'utilisateurs en analysant leur bibliothèque. Les analyses sont alors basées sur une taxonomie qui organise les sujets des ouvrages, cf. Figure 1.
Dans cette configuration, l' *IC* des concepts peut être utilisé pour caractériser le centre d'intérêt d'un utilisateur. Savoir qu'un utilisateur a un manuel de mathématiques sera considéré comme moins informatif que de savoir qu'il a un manuel d'algèbre. En effet, de par la propriété de transitivité de la relation de subsomption, tout livre indexé par *Algèbre* est nécessairement associé à *Maths* et *Sciences*. En d'autres termes, quand la bibliothèque d'un utilisateur sera analysée, tout ouvrage traitant d'*Algèbre* contribuera à renforcer l'idée que l'utilisateur porte de l'intérêt à l'*Algèbre*, aux *Mathématiques* et à la *Science*. Dans les modèles classiques d' *IC* , *l'informativité* du concept *Maths* est simplement vue comme une fonction du nombre de ses instances (i) directes et (ii) indirectes, *i.e*. le cardinal (i) de l'ensemble des livres explicitement annotés par *Maths*, et (ii) de l'ensemble des livres indexés par *Algèbre* ou *Géométrie* dans notre exemple. Par la suite, quand l' *IC* d'un concept est calculé avec l'un des modèles de la littérature, l'information relative au nombre d'occurrences des concepts qui le subsument ne sera pas prise en compte. Par exemple, sur la Figue 1, le nombre d'occurrences du concept *Maths* ne sera pas pris en compte pour estimer l' *IC* du concept *Algèbre*. Toutefois, ce choix de modélisation peut induire des résultats non souhaités et peut être mal adapté à certains contextes d'utilisation. Illustrons le cœur de la problématique sur la base de l'exemple de la bibliothèque. Considérons un cas extrême où un utilisateur a une bibliothèque de 100 livres dont 98 sont explicitement annotés par *Maths*, 1 est annoté par *Algèbre* et 1 par *Antiquité*. Avec un modèle classique, les *ICs* des concepts *Algèbre* et *Antiquité* seront les mêmes. Autrement dit, cela signifie que savoir qu'il y a un livre d'algèbre dans la bibliothèque n'est pas plus informatif que de savoir qu'il y en a un qui traite de l'antiquité, et cela bien que nous sachions que l'utilisateur est extrêmement intéressé par les mathématiques (98 ouvrages). A notre sens, cette remarque met en avant une limite dans l'utilisation des modèles classiques d' *IC* qui ne considèrent qu'une information partielle sur les relations entre les concepts et les instances−ce qui nous parait contre-intuitif dans certains cas d'usage comme dans l'exemple précédent. En effet, dans ce cas, la plupart des gens s'accorderont à affirmer que le sujet *Antiquité* est plus

informatif que celui d'*Algèbre* puisqu'il permet d'identifier un centre d'intérêt possible de l'utilisateur qui ne pouvait jusqu'alors être soupçonné au vu des observations. Cet article propose de définir et d'étudier de nouveaux modèles d'*IC* en considérant que l'estimation de *l'informativité* d'un concept est contextuelle et largement impactée par le fait que nous ayons souvent recours à l'inférence inductive au quotidien. Par exemple, dire que quelqu'un aime les mathématiques tend à renforcer l'idée que ce quelqu'un pourrait bien aimer l'algèbre. Ce point de vue sera désigné tout au long de l'article par *hypothèse d'inférence inductive* (*HII*).

Cet article propose et évalue de nouveaux modèles d'*IC* qui règlent certaines limitations d'usage des estimateurs *d'informativité* de la littérature en implémentant l'*HII*. Le principe consiste à considérer les occurrences de concepts subsumant dans le calcul de l'*IC* du subsumé pour conférer à notre modèle un caractère *intuitif* qui nous semble assez naturel, du moins dans certains cas d'usage des *IC*. Nous nous appuyons sur le cadre théorique des fonctions de croyance pour asseoir notre proposition.

L'article est structuré comme suit. La section 2 introduit le formalisme et dresse un rapide état de l'art des différents modèles d'*IC* existants. La section 3 introduit les notions de la théorie des croyances nécessaires à la compréhension de notre modèle et explique comment les mettre en œuvre pour modéliser l'*HII*. La section 4 est dédiée à l'évaluation et compare notre proposition aux modèles existants dans les usages les plus classiques : l'évaluation est essentiellement orientée autour de l'impact des modèles d'*IC* sur les mesures sémantiques.

2 Formalisme et existant

2.1 Ontologies : formalisme adopté

Considérons une ontologie dont peut être extraite une taxonomie $O = (\preceq, C)$ (en appliquant des règles d'inférence si nécessaire). Cette taxonomie ordonne ($\preceq$) partiellement l'ensemble des concepts C qu'elle définit. $A(c) = \{x \in C \mid c \preceq x\}$ et $D(c) = \{x \in C \mid x \preceq c\}$ représentent alors respectivement l'ensemble des ancêtres et des descendants (non stricts) du concept $c \in C$. L'unique concept sans ancêtre (excepté lui-même) est appelé *root* $(A(root) = \{root\})$. Les concepts n'ayant aucuns descendants exceptés eux-mêmes $(D(c) = \{c\})$ sont appelés des feuilles et sont notés *Leaves*. L'ensemble des feuilles subsumées par le concept $c \in C$ sera noté $Leaves_c$ (*i.e.* $Leaves_c = D(c) \cap Leaves$).

À chaque concept $c \in C$ est attaché un ensemble d'instances et le concept peut alors être considéré comme la classe de ces instances, *e.g.*, le concept *Maths* de la Figure 1 représente l'ensemble des livres qui sont annotés par ce concept. Notons I l'ensemble de toutes les instances (*i.e.*, tous les livres de Lucie dans l'exemple). Pour un concept $c \in C$, deux sous-ensembles de I peuvent lui être associés. Le premier, noté I_c, désigne l'ensemble des instances de I annotées par c (sans considération de la taxonomie). Nous considérerons qu'aucune instance ne peut être annotée simultanément par deux concepts différents dont l'un est descendant de l'autre. Le deuxième, noté $I_{c/\preceq}$, désigne l'ensemble des instances de c obtenu par transitivité de la relation $\preceq$: $I_{c/\preceq} = \bigcup_{x \preceq c} I_x$. Il en découle que $I_{root/\preceq} = I$ et $\forall c \in Leaves$, $I_{c/\preceq} = I_c$.

2.2 Modèles usuels d'IC

La transitivité de la relation $\preceq$ assure que toute instance d'un concept $c \in C$ est une instance de chacun des ancêtres de c (*i.e.* $\forall c_1, c_2 \in C$, $c_1 \preceq c_2 \Rightarrow I_{c_1/\preceq} \subseteq I_{c_2/\preceq}$). Cette notion est généralement utilisée pour décrire la spécificité d'un concept $c \in C$ en comparant le nombre d'instances associées à ce concept $|I_{c/\preceq}|$ relativement à $|I|$. Plus un concept est

restrictif ($|I_{c/\preceq}|$ est petit), plus sa spécificité est grande. Dans la littérature, la spécificité d'un concept est aussi mesurée par la quantité d'information (*IC* : *Information Content*) apportée par le concept. Dans cet article, nous nous référons à la notion d'*IC* définie comme une fonction associant à chaque concept une valeur réelle positive, *i.e.* $IC : C \to \mathbb{R}^+$. Conformément aux règles imposées en modélisation des connaissances, l'*IC* doit décroitre de façon monotone des feuilles à la racine : $\forall c_1, c_2 \in C \;\; c_1 \preceq c_2 \Rightarrow IC(c_1) \geq IC(c_2)$. Ce qui suit décrit les deux principales approches utilisées pour estimer l'*IC* d'un concept.

L'approche intrinsèque est uniquement basée sur la structure topologique de la taxonomie et étudie la position du concept dans cette dernière. Elle ne prend pas en compte les instances associées aux concepts. De nombreux estimateurs ont été proposés basés sur la profondeur du concept, le nombre de ses ancêtres/descendants, le nombre de feuilles qu'il subsume... ou en utilisant une combinaison de ses éléments. Une expression de l'*IC* basée sur un seul de ces éléments est, par exemple, fournie par Seco *et al*. [3] et définie comme étant inversement proportionnelle au nombre de descendants : $IC_{Seco}(c) = 1 - \log|D(c)| / \log|C|$.

Des expressions intrinsèques plus complexes existent, comme par exemple celle de Sanchez *et al*. [4] qui utilise à la fois le nombre de feuilles subsumées par un concept et le nombre de ses ancêtres. De manière générale, les *IC* intrinsèques sont efficaces pour estimer la quantité d'information portée par un concept en analysant les propriétés topologiques de la taxonomie. Cependant, ils ne peuvent pas prendre en compte l'utilisation qui est faite d'un concept dans un cadre applicatif spécifique. Dans de nombreux cas, l'*IC* ne peut pour autant pas être estimé sans tenir compte du contexte dans lequel le concept est utilisé. Par exemple, le concept *Fuzzy logic* est inconnu de la majorité des gens et peut donc être considéré comme très spécifique/informatif dans la plupart des contextes. Cependant, il n'aura que peu d'apport informatif s'il est utilisé pour caractériser les articles publiés dans une revue comme *Fuzzy Sets and Systems*. Afin de surmonter les limitations de l'approche intrinsèque, des éléments d'information extrinsèques (*i.e.* qui ne viennent pas de la taxonomie) peuvent aussi être pris en compte pour estimer l'*IC* du concept.

L'approche extrinsèque est basée sur la théorie de l'information de Shannon où le caractère informatif d'un concept est le résultat de l'analyse statistique d'un corpus de textes. À l'origine, Resnik [2] définit l'*IC* d'un concept comme étant inversement proportionnelle à sa probabilité d'apparition dans le corpus. En considérant que l'information sur l'usage d'un concept puisse être obtenue de l'étude d'un ensemble d'entités annotées par les concepts (*e.g.* livres d'une bibliothèque), la probabilité (notée $p(c)$) qu'une instance appartenant à I soit une instance de c est $p(c) = |I_{c/\preceq}| / |I|$. Le contenu informationnel d'un concept est alors défini par : $IC(c) = -\log p(c)$.

D'autres formulations d'*IC* peuvent être trouvées dans [5]. Les formulations intrinsèques ne prennent pas en compte la manière dont sont utilisés les concepts. En revanche, les formulations extrinsèques n'utilisent qu'une très faible partie de la sémantique exprimée dans la taxonomie : seule l'information portée par les descendants d'un concept est utilisée pour estimer l'*IC*. Par conséquent, aucune de ces deux approches n'implémentent l'*Hypothèse d'Inférence Inductive* (*HII*) : en utilisant ces méthodes, en aucun cas, le nombre d'observations d'un concept ne modifiera notre croyance à observer l'un de ses descendants, *e.g.* savoir qu'il y a 98 livres de *Maths* dans la librairie de Lucie n'influence en aucune manière les *IC* des observations « *Lucie a un livre d'algèbre* » et « *Lucie a un livre d'antiquité* » qui demeurent identiques (dans le cas *intrinsèque* : la taxonomie est indépendante des instances ; dans le cas *extrinsèque* : Lucie a un livre sur chaque sujet). Du fait qu'aucune de ces deux

approches (intrinsèque et extrinsèque) n'est capable de modéliser l'*HII*, les sections suivantes étudient de nouvelles expressions de l'*IC* intégrant cette notion. Notre proposition se base sur des statistiques d'événements imprécis hiérarchisés par la structure de la taxonomie plutôt que d'utiliser la solution d'agréger les deux types d'IC comme proposé dans [17].

3 IC et théorie de l'évidence

Reprenons l'exemple de la section 1 : savoir que *Lucie étudie l'algèbre* est plus informatif que de savoir que *Lucie étudie les mathématiques* – en référence à la Figure 1 où *Algèbre* est une feuille et *Mathématiques* est l'un de ses parents. La seule information *Lucie étudie les mathématiques* signifie que l'on ne peut répondre plus précisément à la question *Qu'étudie Lucie ?* On peut cependant supposer que l'information véhiculée est imprécise par manque de connaissances. Seule une source mieux informée sur Lucie pourrait répondre plus précisément par *Algèbre* ou *Géométrie*. C'est cette remarque que nous allons exploiter pour introduire l'*HII* en utilisant la théorie des fonctions de croyance.

L'idée pour modéliser l'imprécision véhiculée par un concept c d'une taxonomie consiste à lui associer un sous-ensemble Ω_c d'un ensemble fini de labels $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$. La construction de Ω doit respecter les propriétés suivantes : (1) les éléments les plus spécifiques (les feuilles) sont associés à des singletons : $\forall c \in leaves$, $|\Omega_c| = 1$; (2) $\forall c, c' \in C$, $c \prec c'$, alors $\Omega_c \subset \Omega_{c'}$. La première propriété signifie qu'en prenant en compte la taxonomie, une annotation par une feuille est une annotation précise, *i.e.* réduite à un singleton. Ainsi, $\Omega \supseteq \bigcup_{c \in leaves} \{\omega_c\}$ où ω_c est le label associé à $c \in leaves$. La seconde propriété signifie qu'une annotation par un concept c plus spécifique que c' doit être associée à un sous-ensemble de labels plus petit que celui associé à c'. En effet, si on se contentait de construire Ω_c pour des concepts autres que les feuilles comme l'ensemble des labels associés aux concepts feuilles subsumés par c, on pourrait avoir $\Omega_c = \Omega_{c'}$ avec $c \neq c'$, *e.g.* lorsqu'un concept n'a qu'un seul fils. Afin que la propriété (2) soit satisfaite, il est alors nécessaire d'ajouter une feuille fictive à chaque concept qui n'est pas une feuille. A chacune de ces feuilles fictives est associé un nouveau label ω_c ajouté à Ω. ω_c renvoie aux instances de c qui ne font pas implicitement référence aux descendants (exclusifs) de c. Ainsi ω_{Maths} permet de faire référence aux instances du concept *Maths* (livres de Maths) qui ne peuvent pas être considérés comme des livres d'*Algèbre* ou de *Géométrie*. Cela permet de définir Ω_c comme suit : $\Omega_c = \bigcup_{c' \preceq c} \{\omega_{c'}\}$. Dans notre exemple, le concept *Maths* correspond à une réponse imprécise que nous représentons par le sous-ensemble $\Omega_{Maths} = \{\omega_{Algebra}, \omega_{Geometry}, \omega_{Maths}\}$ alors que les réponses *Algèbre* et *Géométrie* sont précises : $\Omega_{Algebra} = \{\omega_{Algebra}\}$ et $\Omega_{Geometry} = \{\omega_{Geometry}\}$.

Cette représentation conduit à introduire une distribution de probabilité sur 2^{Ω} afin de comptabiliser les observations imprécises. Les fonctions définies dans la théorie de l'évidence possèdent les propriétés nécessaires pour propager la *masse* de l'information observée sur les sous-ensembles de Ω. Nous allons maintenant présenter comment elles ont été utilisées afin de modéliser l'*HII*.

3.1 Fonctions de croyance

La théorie des fonctions de croyance a été introduite par Shafer [7] pour modéliser l'imprécision et l'incertitude. Elle a été appliquée dans plusieurs domaines où l'information est fournie par des capteurs imprécis ou des jugements d'experts [8]. Nous résumons dans la suite les principales notions de la théorie des fonctions de croyance.

Soit Ω un ensemble fini. Une fonction de masse *m* est une distribution de probabilité sur

2^{Ω} ($m: 2^{\Omega} \rightarrow [0,1]$), elle est aussi appelée *bpa* (*basic probability assignment*). Soit $A \subseteq \Omega$, $m(A)$ est la proportion de la masse affectée à A lui-même et non à un sous-ensemble plus petit [7]. Dubois *et al*. voient $m(A)$ comme la probabilité qu'un agent ne sache rien de plus que $\omega \in A$ (où ω est la valeur qu'on cherche à connaitre) [9].

Si nous faisons le lien avec la taxonomie et l'ensemble de labels Ω associé, une annotation imprécise véhiculée par un concept $c \in C$ est représentée dans Ω par le sous-ensemble Ω_c. La quantité $m(\Omega_c)$ fait référence à $|I_{c/\preceq}| / |I|$, *i.e.* la probabilité d'observer une instance de *c* et seulement une instance de *c*, ce qui exclut les instances des concepts descendants de *c*.

Les sous-ensembles $A \subseteq \Omega$ tels que $m(A) > 0$ sont appelés éléments focaux et leur ensemble est noté $\mathbb{F}$. Pour mesurer la masse de croyance totale affectée à un sous-ensemble *A*, il faut rajouter à $m(A)$, les quantités $m(B)$ des éléments focaux *B* inclus dans *A*. Cette mesure est modélisée par la fonction de croyance, notée *Bel*, ($Bel: 2^{\Omega} \rightarrow [0,1]$) définie par $Bel(A) = \sum_{B \in \mathbb{F}, B \subseteq A} m(B)$.[1]

Pour la taxonomie, la quantité $Bel(\Omega_c)$ est la somme des masses de tous les éléments focaux inclus dans Ω_c, *i.e.* les événements qui prouvent Ω_c : si une instance d'*Algèbre* est observée, alors une instance de *Maths* est observée. En revanche, avec la fonction *Bel* ou la fonction *p* définie dans [2], observer une instance de *Maths* ne renseigne pas sur ce qu'on peut dire sur *Algèbre*, *i.e.* l'*HII* n'entre pas en jeu. L'approche proposée dans cet article est de modéliser l'*HII* via la fonction de *plausibilité*. La fonction de plausibilité $Pl: 2^{\Omega} \rightarrow [0,1]$ est définie comme suit : $Pl(A) = \sum_{B \in \mathbb{F}, B \cap A \neq \varnothing} m(B)$ où $Pl(A)$ exprime le degré auquel on estime que l'évènement A est plausible. Par exemple, puisque $\Omega_{Algèbre} \subset \Omega_{Maths}$, si $\Omega_{Maths} \in \mathbb{F}$ alors la plausibilité de $\Omega_{Algèbre}$ est renforcée. Ainsi, plus la probabilité d'observer le concept *Maths* est grande, plus la plausibilité que l'annotation soit *Algèbre* est renforcée. Quand les observations (éléments focaux) sont imprécises, la probabilité d'un événement *A*, notée *Pr(A)*, est aussi imprécise et est encadrée comme suit : $Pr(A) \in [Bel(A), Pl(A)]$.

Dans la théorie de l'évidence, la paire (Bel, Pl) est la contrepartie de la probabilité traditionnelle Pr. L'utilisation judicieuse de cette paire en lieu et place de *Pr* serait de conserver l'encadrement. Cependant, comme l'*IC* d'un concept est communément entendu comme une valeur dans $\mathbb{R}^+$, un choix doit être effectué. De plus, d'autres fonctions que *Bel* et *Pl* sont proposées dans le cadre de la théorie de l'évidence. Par exemple, la mesure de probabilité pignistique *BetP* de Smets [10] est calculée à partir de la fonction de masse *m*. La mesure *BetP* est définie via sa distribution de probabilité associée notée $betP_m$ comme suit : $\forall \omega \in \Omega, \quad betP_m(\omega) = \sum_{A \in \mathbb{F}, A \ni \omega} m(A) / |A|$.

Pour $\omega \in \Omega$, la distribution de probabilité $betP_m$ peut être vue comme la moyenne pondérée de probabilités uniformes sur les éléments focaux contenant ω où les poids sont les masses de ces éléments focaux [9]. Comme la fonction de plausibilité, la mesure de probabilité pignistique tire profit d'informations sur les ancêtres pour inférer de l'information sur les descendants.

3.2 Propagation de l'information

Revenons sur l'exemple de l'utilisateur qui possède une bibliothèque de 100 livres : 98

---

[1] *Pour un concept c, $Bel(\Omega_c)$ correspond à la quantité définie par Resnik [2] comme la probabilité d'observer une instance de c, noté $p(c)$: $Bel(\Omega_c) = p(c) = |I_{c/\preceq}| / |I|$. Notons que p ne devrait pas être vue comme une distribution de probabilité sur C [2] vu que $p(root) = 1$, $root \in C$ et les éléments de C ne sont pas considérés indépendants.*

sont explicitement annotés par *Maths*, 1 annoté par *Algèbre* et 1 annoté par *Antiquité*. Dans ce cas, les concepts observés sont associés aux éléments focaux suivants : $\mathbb{F}=\{\Omega_{Maths}, \Omega_{Algèbre}, \Omega_{Antiquité}\}$. La Figure 3 montre les masses associées à cette partie de la taxonomie pour cet exemple.

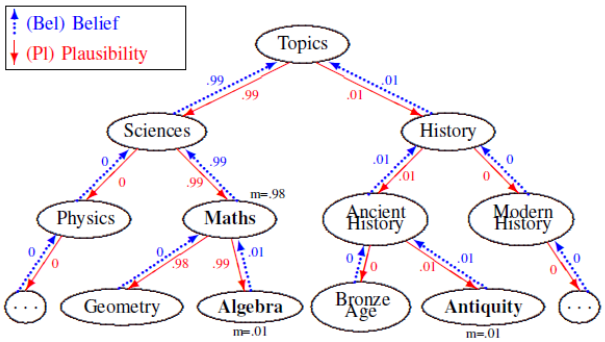

Figure 3 - Propagation des masses par *Bel* et *Pl* dans une taxonomie

Comme le montre la Figure 3, la fonction de croyance *Bel* correspond à une propagation des instances du bas vers le haut : si la bibliothèque contient des livres d'algèbre, cette bibliothèque contient *donc* des livres de mathématiques (au moins des livres d'algèbre). Ceci correspond au mécanisme classique utilisé dans les approches probabilistes (cf. Resnik [2]). En revanche, la fonction de plausibilité *Pl* permet de propager les masses dans les deux sens, et notamment des ancêtres vers les descendants, l'observation d'un ancêtre rend plausible (crédible) l'observation de chacun de ses descendants : avoir des témoignages sur la présence de livres de mathématiques dans la bibliothèque ne prouve pas qu'il y ait des livres de géométrie, mais renforce néanmoins la plausibilité de cet événement. Cette *intuition* correspond à l'*HII*.

3.3 *IC* et fonctions de croyance

Soient une fonction de masse *m*, *Bel* et *Pl* respectivement les fonctions de croyance et de plausibilité associées à *m*, Ω l'ensemble des labels associés à *C*. Plusieurs propositions de nouveaux *IC* qui intègrent la propagation d'instances du haut vers le bas sont alors envisageables avec les fonctions de la théorie des croyances introduites. Ainsi, pour tout $c \in C$, on peut poser :

a- $IC_{Pl}(c) = -\log Pl(\Omega_c)$;

b- $IC_{BetP_m}(c) = -\log \sum_{\omega \in \Omega_c} betP_m(\omega)$

Notons que l'*IC* défini par la fonction *Bel*, *i.e.*, $IC_{Bel}(c) = -\log Bel(\Omega_c)$, ne permet pas de propager les masses des ancêtres vers les descendants et que par ailleurs, la contrainte de monotonie est vérifiée pour chacun des *IC* proposés. Le choix de la fonction logarithmique découle de l'application de la définition classique du contenu informationnel et fait donc référence à la théorie de l'information de Shannon.

4 Évaluations

Notre principale contribution porte sur la modélisation de l'*HII* dans l'IC_{Pl} et l'IC_{BetP_m}. Cette proposition répond aux limites des *IC* existants (c.f. Section I) en permettant une estimation plus fine de *l'informativité* d'un concept en considération de l'*HII*. Discuter la pertinence de considérer cette hypothèse en fonction d'un contexte particulier dépasse le cadre de cette contribution. L'idéal serait cependant de tester la performance des *IC* proposés dans la tâche de modélisation de l'*HII*. Une telle évaluation est cependant difficile à mettre en œuvre car il faudrait au préalable définir une procédure (*e.g.* mesure) permettant de comparer des modèles implémentant cette hypothèse. Néanmoins, puisque (i) les applications et les apports de notre contribution ont été soulignés, (ii) aucun modèle implémentant l'*HII* n'a été proposé jusque-là à notre connaissance, et (iii) aucun test ne permet d'évaluer la performance avec laquelle un *IC* modélise cette hypothèse, nous proposons d'évaluer notre contribution indirectement, en discutant l'impact des *IC* sur

la performance des systèmes qui reposent sur leur utilisation. Nous souhaitons en particulier nous assurer que les *IC* proposés n'impactent pas négativement la performance de ces systèmes. Pour cela nous proposons de discuter l'impact des *IC* sur la performance des mesures de similarité sémantique. Tous les résultats présentés peuvent être reproduits en utilisant le code source et les données publiés à l'adresse : https://github.com/sharispe/published_xp.

De nombreuses mesures permettent d'estimer la similarité sémantique de deux concepts définis dans une taxonomie ; celles-ci sont largement utilisées en RI, TALN et dans différentes techniques de raisonnement approché. Parmi les plus performantes, plusieurs mesures reposent sur les modèles classiques d'*IC* introduits en section II. Nous présentons deux mesures populaires que nous utiliserons dans nos évaluations. Resnik [2] définit la similarité des concepts c_1 et c_2 comme l'*IC* de leur ancêtre commun le plus informatif :

$$sim_{Resnik}(c_1, c_2) = \max_{a \in A(c_1) \cap A(c_2)} IC(a)$$

L'adaptation proposée par Lin [11] considère aussi la spécificité des concepts comparés :

$$sim_{Lin}(c_1, c_2) = \frac{sim_{Resnik}(c_1, c_2)}{IC(c_1) + IC(c_2)}$$

Les mesures de similarité sémantique sont classiquement évaluées au regard de leur capacité à mimer l'appréciation qu'a l'homme de la similarité sémantique. La performance des mesures est alors évaluée par l'étude des corrélations de Pearson/Spearman entre des scores attendus et estimés. Les jeux de tests se composent de paires de concepts associées à des scores attendus de similarité, généralement la moyenne des similarités exprimées par différents participants. Nous proposons ici de comparer la performance des mesures utilisant des modèles d'*IC* classiques et ceux proposés dans cet article. Trois jeux de tests ont été utilisés : (i) Rubenstein & Goodenough (RG) [12], 65 paires de mots, (ii) Miller & Charles (MC) [13], 28 paires de concepts et (iii) SimLex999 (SL) [14], pour lequel nous avons utilisé les 666 paires de noms qu'il contient.

L'ordre partiel des concepts comparés dans notre évaluation est fourni par WordNet 3.1 [15] – seule la partie associée aux noms a été considérée. Les jeux de tests utilisés fournissent des paires de concepts et les scores de similarité associés. Le calcul de l'*IC* de Resnik et des *IC* proposés dépendant de l'observation de fréquences d'utilisation des concepts, nous avons utilisé les statistiques fournies par le Princeton WordNet Gloss Corpus (http://wordnet.princeton.edu/glosstag.shtml). Les expériences menées reposent sur les mesures et les *IC* implémentés dans la Semantic Measures Libary (http://www.semantic-measures-library.org) [16]. Les *IC* proposés dans cet article ont eux aussi été développés en utilisant cette librairie.

Le Tableau 1 présente les corrélations de Spearman et Pearson obtenues avec chaque configuration de mesure pour les différents jeux de test (RG, MC et SL). Les résultats sont présentés pour les *IC* proposés (Belief, Pignistic et Plausibility) et différents *IC* (voir section 2). Nous avons aussi considéré une formulation intrinsèque de l'*IC* proposé par Resnik, ici dénommée Resnik (i) pour intrinsèque. Nous rappelons que les *IC* basés sur la plausibilité et sur la probabilité pignistique implémentent l'*HII*, ils sont associés au symbole * dans le tableau 1.
Les résultats soulignent que les modèles d'*IC* proposés (Belief, Pignistic, Plausibility) rivalisent avec les meilleurs modèles d'*IC* lorsqu'ils sont évalués au travers de l'étude des performances des mesures sémantiques. Ce résultat n'a rien de surprenant pour le modèle basé sur la fonction de croyance car celui-ci n'est au final qu'une variante de l'*IC* proposé par Resnik. Cependant, de façon intéressante, ces résultats soulignent qu'en plus d'implémenter l'*hypothèse d'inférence inductive*, les *IC* reposant sur la plausibilité et la probabilité pignistique permettent d'obtenir des performances comparables à celles des meilleurs modèles d'*IC*. A noter que la faible

performance des mesures reposant sur l'*IC* Resnik (i) montre que les résultats observés ne découlent pas du fait que l'*IC* n'ait pas d'impact sur la performance des mesures. Nous avons aussi analysé les corrélations entre les estimations d'*IC* obtenues à l'aide des différents modèles. Les résultats obtenus mettent en évidence, comme souhaité, que les *IC* implémentant l'*HII* se comportent différemment. Cela souligne alors que ces modèles sont performants – au minimum dans un contexte de calcul de similarité sémantique. Nous avons aussi observé que l'*IC* basé sur la probabilité pignistique a un comportement qui se rapproche plus de l'*IC* belief que de l'*IC* plausibilité. Conformément avec les inégalités $\forall c \in C, IC_{Pl}(c) \leq IC_{BetPm}(c) \leq IC_{Bel}(c)$, ces résultats suggèrent que l'*IC* basé sur la probabilité pignistique permet d'obtenir un compromis intéressant, *i.e.* de modéliser l'*HII* tout en s'assurant d'obtenir des estimations d'*IC* qui ne soient pas radicalement différentes de celles obtenues traditionnellement.

Tableau 1 – Performance des mesures

	Pearson		**Spearman**	
	Resnik	Lin	Resnik	Lin
	Rubenstein & Goodenough			
IC Belief	0.478	0.478	0.455	0.432
IC Pignistic*	0.481	0.480	0.455	0.424
IC Plausibility*	0.498	0.498	0.454	0.423
IC Resnik	0.477	0.477	0.468	0.439
IC Resnik (i)	0.339	0.339	0.320	0.320
IC Seco	0.482	0.480	0.455	0.443
IC Sanchez	0.516	0.514	0.451	0.435
	Miller & Charles			
IC Belief	0.761	0.833	0.756	0.797
IC Pignistic*	0.774	0.843	0.758	0.794
IC Plausibility*	0.827	0.836	0.769	0.791
IC Resnik	0.809	0.838	0.793	0.804
IC Resnik (i)	0.280	0.281	0.266	0.266
IC Seco	0.808	0.841	0.760	0.808
IC Sanchez	0.836	0.847	0.775	0.765
	SimLex 666			
IC Belief	0.528	0.597	0.531	0.582
IC Pignistic*	0.534	0.594	0.533	0.583
IC Plausibility*	0.527	0.564	0.521	0.565
IC Resnik	0.114	0.114	0.108	0.108
IC Resnik (i)	0.538	0.601	0.527	0.588
IC Seco	0.482	0.480	0.525	0.592
IC Sanchez	0.541	0.583	0.527	0.583

Conclusion

Nous avons présenté de nouveaux modèles d'*IC* basés sur les fonctions de croyance. Ils permettent d'estimer le contenu informationnel de concepts définis dans une taxonomie en tenant compte à la fois de l'ordre partiel et de statistiques relatives à leur usage (*e.g.* dans des textes). En particulier, par la définition de deux *IC* extrinsèques basés sur la plausibilité et sur la probabilité pignistique, nous avons introduit deux modèles d'IC implémentant l'hypothèse d'inférence inductive (*HII*), *i.e.* que les observations d'un concept impactent (i) *l'informativité* à la fois du concept en question et de ses généralisations, mais aussi (ii) *l'informativité* des concepts qu'il généralise. La motivation concernant l'utilisation d'un tel modèle d'*évocation,* est qu'intuitivement un fait augmente l'état de croyance des faits qu'il généralise, *e.g.* apprendre qu'une personne apprécie les livres de *Maths* peut nous conforter dans l'idée qu'elle puisse aimer les livres d'*Algèbre*. Ces modèles ont la propriété intéressante de répondre à l'incapacité des modèles d'*IC* existants à modéliser ce comportement – cela malgré le rôle important que celui-ci semble avoir dans la mise en place de processus cognitifs. De plus, en considérant certains concepts comme des expressions imprécises d'autres concepts, ces nouveaux modèles d'*IC* apportent un regard original sur l'information portée par l'observation de concepts en considération d'un ordre partiel les structurant. Les premières évaluations empiriques de ces mesures, basées sur l'étude de l'impact des *IC* sur la performance des mesures sémantiques, montrent que les modèles proposés rivalisent avec les modèles classiques les plus performants. Des résultats complémentaires sur les corrélations entre ICs sont proposés dans [18]. Pour toutes ces

raisons, nous pensons que les modèles d'*IC* implémentant l'*HII* sont adaptés à de nombreux contextes applicatifs, *e.g.* recommandation, RI, TALN. Néanmoins, nous souhaitons souligner que la sémantique associée au modèle implémentant l'*HII* n'est probablement pas adaptée à tous les contextes – des analyses supplémentaires dans différents contextes applicatifs méritent d'être menées pour compléter nos résultats et affiner notre compréhension des modèles d'*IC*. De façon intéressante, notre étude a permis de faire le lien entre des contributions de domaines de recherche qui avaient jusque-là peu de connections. En effet, en soulignant le lien étroit entre la fonction de croyance *Bel* et la probabilité à la base de la définition de l'*IC* de Resnik, ainsi que l'intérêt de considérer la plausibilité et la probabilité pignistique pour implémenter l'*HII*, nous avons mis en évidence l'impact intéressant que le cadre des fonctions de croyance pourrait avoir sur les travaux de recherche portant sur l'estimation d'IC, les mesures sémantiques et plus généralement, la recherche approchée basée sur l'utilisation d'ontologies.

Références

[1] S. Staab and R. Studer, Handbook on ontologies. Springer Science & Business Media, 2010.

[2] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," in Proceedings of the 14th International Joint Confer- ence on Artificial Intelligence IJCAI, vol. 1, 1995, pp. 448–453.

[3] N. Seco, T. Veale, and J. Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet," in 16th European Conference on Artificial Intelligence. IOS Press, 2004, pp. 1–5.

[4] D. Sanchez, M. Batet, and D. Isern, "Ontology-based information content computation", Knowledge-Based Systems, vol. 24, no. 2, pp.297–303, Mar. 2011.

[5] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis," ArXiv, vol. 1310.1285, p. 140, Oct. 2013.

[6] S. Harispe, D. Sanchez, S. Ranwez, S. Janaqi, and J. Montmain, "A Framework for Unifying Ontology-based Semantic Similarity Measures: a Study in the Biomedical Domain," Journal of Biomedical Informatics, vol. 48, pp. 38–53, 2013.

[7] G. Shafer, A mathematical theory of evidence (Vol. 1). Princeton: Princeton university press, 1976.

[8] A. Imoussaten, J. Montmain, and G. Mauris, "A multicriteria deci- sion support system using a possibility representation for managing inconsistent assessments of experts involved in emergency situations," International Journal of Intelligent Systems, vol. 29, no. 1, pp. 50–83, 2014.

[9] D. Dubois and H. Prade, "Formal representations of uncertainty," Decision-Making Process: Concepts and Methods, pp. 85–156, 2009.

[10] P. Smets and R. Kennes, "The transferable belief model," Artificial intelligence, vol. 66, no. 2, pp. 191–234, 1994.

[11] D. Lin, "An Information-Theoretic Definition of Similarity," in 15th International Conference of Machine Learning, Madison,WI, 1998, pp.296–304.

[12] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," Communications of the ACM, vol. 8, no. 10, pp. 627–633, Oct. 1965.

[13] G. A. Miller and W. G. Charles, "Contextual Correlates of Semantic Similarity," Language & Cognitive Processes, vol. 6, no. 1, pp.1–28, 1991.

[14] F. Hill, R. Reichart, and A. Korhonen, "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation," Aug. 2014. [Online]. Available: http://arxiv.org/abs/1408.3456

[15] G. A. Miller, "WordNet: a lexical database for English," Communica- tions of the ACM, vol. 38, no. 11, pp. 39–41, 1998.

[16] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain, "The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies," Bioinformatics, vol. 30, no. 5, pp. 740–742, 2014.

[17] W. Xu. Modélisation et exploitation de base de connaissances dans le cadre du web des objets, Thèse de l'Université Pierre et Marie Curie, 2015.

[18] Harispe, S., Imoussaten, A., Trousset, F., Montmain, J. (2015). On the consideration of a Bring-to-mind Model for Computing the Information Content of Concepts defined into Ontologies, FUZZ IEEE, Istanbul, Turkey.