Comparaison et Évaluation de Mesures de Similarité entre Concepts d'un Treillis

Florent Domenach*, George Portides**

* Akita International University, Yuwa, Akita-city 010-1292, Japan
fdomenach@aiu.ac.jp,
** University of Nicosia, 46 Makedonitissas Ave., PO Box 24005, 1700 Nicosia, Cyprus

Résumé. Cet article se situe dans le cadre de l'analyse de concepts formels (ACF) qui fournit des classes (les extensions) d'objets partageant des caractères similaires (les intensions), une description par des attributs étant associée à chaque classe. Dans un article récent, une nouvelle mesure de similarité entre deux concepts dans un treillis de concepts a été introduite, permettant une normalisation par la taille du treillis. Dans cet article, nous comparons cette mesure de similarité avec des mesures existantes, soit basées sur la cardinalité des ensembles ou issues de la conception d'ontologies et basées sur la structure hiérarchique du treillis. Une comparaison statistique avec des méthodes existantes est effectuée et testée pour leur consistance.

1 Introduction

Cet article est un résumé de Domenach et Portides (2016). Les mesures de similarité on été largement utilisées, en particulier dans le domaine biomédical (Nguyen et Al-Mubaid, 2006) ou dans le web sémantique pour le traitement du langage naturel (Seco et al., 2004). Cependant, la plupart de ces applications reposent sur une structure ontologique arborescente pour quantifier le degré de similarité de deux concepts. Le but de cet article est d'étendre ces mesures de similarité au cadre plus général fourni par les treillis et l'analyse de concepts formels, et d'évaluer et de comparer la mesure introduite dans Domenach (2015).

2 Définitions

Analyse de Concepts Formels Nous rappelons ici les notations standards utilisées en analyse de concepts formels (ACF) et nous renvoyons le lecteur à Ganter et Wille (1999) pour plus de détails. Un *contexte formel* (O, A, I) est défini comme un ensemble O d'objets, un ensemble A d'attributs, et une relation binaire $I \subseteq O \times A$. $(o, a) \in I$ signifie que "l'objet o est lié à l'attribut a par la relation I". Deux opérateurs de dérivation peuvent être définis sur les ensembles d'objets et d'attributs comme suit, $\forall O_1 \subseteq O, A_1 \subseteq A : O_1' = \{a \in A : \forall o \in O_1, (o, a) \in I\}$, $A_1' = \{o \in O : \forall a \in A_1, (o, a) \in I\}$. Les deux opérateurs $(\cdot)'$ définissent une correspondance de Galois entre l'ensemble des parties des objets $\mathcal{P}(O)$ et l'ensemble des parties des attributs

$\mathcal{P}(A)$. Une paire $(O_1, A_1), O_1 \subseteq O, A_1 \subseteq A$, est un *concept formel* ssi $O'_1 = A_1$ et $A'_1 = O_1$. O_1 est appelé *l'extension* et A_1 *l'intension* du concept.

L'ensemble de tous les concepts formels, ordonnés par inclusion d'extensions (ou dualement par l'inclusion d'intensions), *i.e.* $(O_1, A_1) \leq (O_2, A_2)$ ssi $O_1 \subseteq O_2$ (ou dualement $A_2 \subseteq A_1$), forme un treillis complet (Barbut et Monjardet, 1970), appelé *treillis de concepts* ou treillis de Galois.

Relation d'Emboîtement Un aspect intéressant des treillis de concepts est les nombreux cryptomorphismes équivalents existants (Caspard et Monjardet, 2003). La relation d'emboîtement (Domenach et Leclerc, 2004) associée au treillis de concepts est une relation binaire $\mathcal{O}$ sur $\mathcal{P}(A)$ telle que $(X, Y) \in \mathcal{O} \iff X \subset Y$ et $X'' \subset Y''$. En d'autres termes, deux ensembles sont emboîtés si l'un est un sous-ensemble de l'autre et s'ils ont une fermeture différente.

3 Mesures de Similarité Existantes entre Concepts

De nombreuses mesures de similarité existantes peuvent être adaptées aux treillis de concepts. Elles peuvent être divisées en trois catégories principales : la première, basée sur le modèle de Tversky (1977), ne prend en compte les concepts que comme des ensembles, ici d'attributs, afin de calculer la similarité entre deux concepts. La seconde, prenant son origine dans les études d'ontologies, utilise le diagramme de Hasse associé au treillis de Galois pour évaluer les distances entre concepts. Enfin, la troisième catégorie concerne les mesures de similarité sémantiques utilisant le contenu de l'information.

Similarités Ensemblistes. Les mesures de similarité basées sur des ensembles peuvent être exprimées en utilisant le modèle de similitude de Tversky. It est défini comme suit : étant donné deux concepts $C_1 = (O_1, A_1)$ et $C_2 = (O_2, A_2)$, avec $\alpha, \beta \geq 0$,

$$S(C_1, C_2) = \frac{|O_1 \cap O_2|}{|O_1 \cap O_2| + \alpha|O_1 - O_2| + \beta|O_2 - O_1|}$$

Suivant les valeurs de α et β, l'indice de Tversky peut être vu comme une généralisation de l'indice de Jaccard (1901) ($\alpha = 1, \beta = 1$), le coefficient de Dice (1945) ($\alpha = \beta = 1/2$) ou la mesure d'inclusion ($\alpha = 0, \beta = 1$).

Similarités Ontologiques. Les mesures suivantes de similarité sont inspirées par le vaste corpus de travaux existants sur les ontologies en logique de description (`DL`), *i.e.*, basées sur un ensemble de concepts, relations et individus représentés dans une `DL`. Nous supposons simplement que le plus proche ancêtre commun (least common subsumer `lcs`) de deux concepts existe, à condition qu'il n'y ai pas de cycle dans les définitions des concepts (Baader et al., 1999).

La structure hiérarchique du treillis est utilisée pour calculer la similarité entre deux concepts C_1 et C_2, en ne considérant que les liens taxonomiques de l'ontologie et le treillis $\mathbb{L}$ comme une généralisation d'un arbre. Afin de définir les similarités dans ce cadre, nous devons définir le `lcs` : $\texttt{lcs} = \texttt{lcs}(C_1, C_2) = C_1 \wedge C_2$, la longueur $length(C_1, C_2)$ comme la

TAB. 1 – *Similarités ontologiques*

Rada et al. (1989)	Rada	$= \frac{1}{length(C_1,C_2)+1}$
Wu et Palmer (1994)	WuPa	$= \frac{2*depth(\texttt{lcs})}{depth(C_1)+depth(C_2)}$
Leacock et Chodorow (1998)	LeCh	$= -\log(\frac{length(C_1,C_2)+1}{2\times depth(L)})$
Pekar et Staab (2002)	PeSt	$= \frac{depth(\texttt{lcs})}{length(C_1,\texttt{lcs})+length(C_2,\texttt{lcs})+depth(\texttt{lcs})}$
Zhong et al. (2002)	Zho	$= 1 - (\frac{1}{2^{depth(C_1)+1}} + \frac{1}{2^{depth(C_2)+1}} - \frac{1}{2^{depth(\texttt{lcs})}})$
Nguyen et Al-Mubaid (2006)	NgAl	$= \log(2 + (length(C_1, C_2) - 1)\times (depth(L) - depth(\texttt{lcs})))$

TAB. 2 – *Fonctions IC*

Resnik (1995)	Res	$= \frac{\lvert O_1\rvert}{\lvert O\rvert}$
Seco et al. (2004)	Seco	$= 1 - \frac{\log(hypo(C))}{\log(\lvert L\rvert-1)}$
Zhou et al. (2008)	Zhou	$= k \times (1 - \frac{\log(hypo(C))}{\log(\lvert L\rvert-1)}) + (1-k) \times \frac{\log(depth(C)+1)}{\log(depth(L)+1)}$
Sánchez et al. (2011)	San	$= -\log(\frac{\frac{leaves(C)}{hypo(C)}+1}{numberofleaves+1})$

distance topologique dans le diagramme de Hasse du treillis et $depth(C_1) = length(C_1, 0_L)$ comme la profondeur du concept C_1, i.e. la distance entre C_1 et le concept minimal de L. La profondeur du treillis est $depth(L) = max_{x\in L}(depth(x))$. Les similarités ontologiques peuvent être trouvées dans le tableau 1.

Similarités basées sur le Contenu Informatif. Une autre approche, en particulier utilisée dans l'étude de similarités sémantiques entre les mots dans Wordnet, améliore les mesures précédentes en augmentant les concepts avec leur Contenu Informatif (Information Content, IC) dérivé d'étiquettes de sens basé sur le corpus ou sur le corpus brut non annoté (Resnik, 1995). Nous pouvons appliquer une approche similaire dans notre cadre de treillis de concepts en définissant d'abord la notion d'IC dans le cadre de l'ACF de la manière suivante : l'IC d'un concept fournit une estimation de son degré de généralité / spécificité, et est une fonction croissante, i.e. a est hyperonyme de $b \Rightarrow IC(a) < IC(b)$. IC est une mesure de la spécificité pour un concept, des valeurs élevées sont associées à des concepts plus spécifiques, tandis que des valeurs inférieures sont plus générales. Les différentes fonctions IC dans le tableau 2 capturent différents aspects du IC, où $hypo(C)$ ($leaves(C)$) est nombre de concepts (co-atomes) au-dessus de C. Chaque similarité du tableau 3 a été implémentée en utilisant chacune des fonctions IC du tableau 2.

TAB. 3 – *Similarités basées sur IC*

Resnik (1995)	Res	$= IC(\texttt{lcs})$
Jiang et Conrath (1997)	JC	$= \frac{1}{IC(C_1)+IC(C_2)-2\times IC(\texttt{lcs})}$
Lin (1998)	Lin	$= \frac{2\times IC(\texttt{lcs})}{IC(C_1)+IC(C_2)}$

Similarités basées sur la Relation d'Emboîtement. Soit C un concept du treillis de Galois $\mathbb{L}$, et définissons $o(C)$ comme l'ensemble des attributs est emboîtés avec C : $o(C) = \{k \in M : (C, C \cup \{k\}) \in \mathcal{O}\}$. $o(C)$ est l'ensemble des attributs qui, quand ajoutés à l'intension de C, crée un concept différent. On peut alors définir (Domenach, 2015) une mesure de similarité : $\forall C_1, C_2 \in \mathbb{L}$, $Over(C_1, C_2) = \frac{|o(C_1 \wedge C_2)|}{|o(C_1) \cup o(C_2)|}$. Cette mesure est basée sur l'idée de prendre en compte la largeur du treillis. Deux concepts vont être plus similaires si ils sont proches dans le treillis et s'ils ne partagent pas d'attributs avec d'autres concepts. C'est une mesure de similarité de deux concepts en relation avec tous les autres concepts.

4 Experimentations

La mesure proposée a été étudiée grâce à une simulation en C♯, où nous avons généré aléatoirement des tables booléennes 20×20, avec des densités variant de 20% à 40 %, avec le treillis associé. Nous avons choisi au hasard 2 concepts différents et calculé 22 similarités entre eux. Cette simulation a été répété 1000 fois.

Un regroupement hiérarchique utilisant un lien moyen a été effectué sur les corrélations de Pearson entre les différentes similarités (dendrogramme de la figure 1). Pearson a été utilisé car nos mesures de similarités ne sont pas normalisées. Il y a deux classes évidentes, mais pas de caractéristiques claires ressortent. Cependant, les similarités sont dans une certaine mesure regroupées selon la classification de la section 3, voir par exemple les similarités ensemblistes.

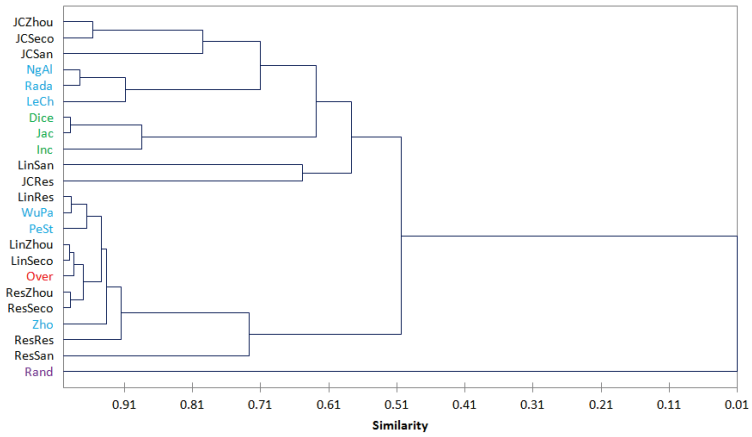

FIG. 1 – *Regroupement hiérarchique utilisant un liant moyen de la matrice de corrélation de toutes les similarités.*

Les corrélations de Pearson de la mesure de similarité d'emboîtement avec les autres mesures, avec sa signifiance, est présentée dans le tableau 4. Il est clair que cette mesure est significativement différente de la mesure aléatoire de référence (Rand). Les estimations des corrélation restantes doivent toutefois être prises avec prudence, car la linéarité entre les mesures mentionnées ci-dessus est encore à étudier.

TAB. 4 – *Corrélation de Pearson et l'importance de la mesure de similarité d'emboîtement avec les 22 autres mesures.*

Jac	Dice	Inc	LeCh	PeSt	Rada	WuPa	LinRes	ResRes	JCRes	LinSec
.444	.422	.389	.588	.917	.575	.950	.934	.893	.258	.982
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
ResSec	JCSec	LinZho	ResZho	JCZho	NgAl	Zho	LinSan	ResSan	JCSan	Rand
.967	.613	.988	.976	.680	.680	.949	.242	.707	.296	.014
.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.650

5 Conclusion

L'implémentation et l'analyse résultante a permis une comparaison de la similarité d'emboîtement avec d'autres mesures. Il y a plusieurs corrélations fortes en particulier avec les similitudes de Lin et de Resnik combinées avec les fonctions IC de Zhou et de Seco, ces cinq mesures (LinZhou, LinSeco, Over, ResZhou, ResSeco) formant un groupe clair. Cela peut être interprété comme capturant des informations similaires utilisant uniquement la structure du treillis, liant l'IC basée sur le nombre de concepts majorant avec la largeur du treillis comme définie par la relation d'emboîtement. Une analyse plus approfondie reste à mener sur les complexités de calcul des différentes mesures et sur leurs relations possibles. Enfin, nous prévoyons d'étudier la robustesse statistique afin de minimiser / éliminer l'effet des valeurs influentes, ainsi que de faire une telle comparaison sur des ensembles de données réels.

Références

Baader, F., R. Küsters, et R. Molitor (1999). Computing least common subsumers in description logics with existential restrictions. In *IJCAI*, Volume 99, pp. 96–101.

Barbut, M. et B. Monjardet (1970). *Ordres et classification : Algèbre et combinatoire (tome II)*. Paris : Hachette.

Caspard, N. et B. Monjardet (2003). The lattices of moore families and closure operators on a finite set : a survey. *Disc. App. Math. 127*, 241–269.

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

Domenach, F. (2015). Similarity measures of concept lattices. In B. Lausen, S. Krolak-Schwerdt, et M. Bohmer (Eds.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 89–99.

Domenach, F. et B. Leclerc (2004). Closure systems, implicational systems, overhanging relations and the case of hierarchical classification. *Math. Soc. Sci. 47 (3)*, 349–366.

Domenach, F. et G. Portides (2016). Similarity measures on concept lattices. In A. Wilhelm et H. KestlerLausen (Eds.), *Analysis of Large and Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 159–169. Springer International Publishing.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Mathematical Foundations*. Springer.

Jaccard, P. (1901). étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles 37*, 547–579.

Jiang, J. J. et D. W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Leacock, C. et M. Chodorow (1998). Combining local context and wordnet similarity for word sense identification. *WordNet : An electronic lexical database 49*(2), 265–283.

Lin, D. (1998). An information-theoretic definition of similarity. In *ICML*, Volume 98, pp. 296–304.

Nguyen, H. A. et H. Al-Mubaid (2006). New ontology-based semantic similarity measure for the biomedical domain. In *Granular Computing, 2006 IEEE International Conference on*, pp. 623–628. IEEE.

Pekar, V. et S. Staab (2002). Taxonomy learning : factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1–7. Association for Computational Linguistics.

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

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Sánchez, D., M. Batet, et D. Isern (2011). Ontology-based information content computation. *Knowledge-Based Systems 24*(2), 297–303.

Seco, N., T. Veale, et J. Hayes (2004). An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, Volume 16, pp. 1089.

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

Wu, Z. et M. Palmer (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138. Association for Computational Linguistics.

Zhong, J., H. Zhu, J. Li, et Y. Yu (2002). Conceptual graph matching for semantic search. In *Conceptual structures : Integration and interfaces*, pp. 92–106. Springer.

Zhou, Z., Y. Wang, et J. Gu (2008). A new model of information content for semantic similarity in wordnet. In *Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on*, Volume 3, pp. 85–89. IEEE.

Summary

This paper falls within the framework of Formal Concept Analysis which provides classes (the extents) of objects sharing similar characters (the intents), a description by attributes being associated to each class. In a recent paper by the first author, a new similarity measure between two concepts in a concept lattice was introduced, allowing for a normalization depending on the size of the lattice. In this paper, we compare this similarity measure with existing measures, either based on cardinality of sets or originating from ontology design and based on the graph structure of the lattice. A statistical comparison with the existing methods is carried out, and the output of the measure is tested for consistency.