Mesure de Similarité entre Treillis Basée sur des Correspondances Explicites

Florent Domenach*

* Akita International University, Yuwa, Akita-city 010-1292, Japan
fdomenach@aiu.ac.jp,

Résumé. Ce document se situe dans le cadre de l'analyse de concepts formels (ACF), une méthode de hiérarchisation algébrique des données basée sur la notion d'intension / extension, partageant maximalement attributs et objets. Nous présentons ici une mesure de similarité basée sur des correspondances entre deux treillis de Galois, définie par un modèle expressif utilisant des correspondances entre objets et entre attributs des deux treillis. Un point clé de notre approche est que ces correspondances peuvent ne pas être des fonctions, associant un objet (resp. attribut) d'un treillis avec plusieurs objets (resp. attributs) de l'autre treillis.

1 Introduction

Cet article est un résumé de Domenach et Rajabi (2015). Les treillis sont des objets polymorphes : comme ensembles ordonnés, les treillis sont une généralisation naturelle de différentes structures comme des arbres, arbres faibles ou pyramides. Ils forment également la structure sous-jacente de l'analyse des concepts formels articulant la dualité entre l'intention et l'extension. En tant que tel, ils peuvent être considérés comme une méthode algébrique de hiérarchisation des données basées sur des attributs et des objets maximalement partagés.

Le problème considéré ici est la quantification de la similarité entre deux treillis donnés, pouvant être définis sur ensembles d'objets et d'attributs différents. Idéalement, une telle mesure devrait prendre des valeurs élevées pour des treillis similaires et des valeurs faibles pour des treillis très dissemblables. La figure 1 montre deux exemples de treillis - la question est de savoir si ces treillis sont "proches", et à quel point ? L'ACF étant particulièrement utilisée dans la recherche d'information et de représentation des connaissances, étudier les mesures de similarité est particulièrement pertinent pour la comparaison des treillis. Nous renvoyons le lecteur à Domenach (2015) où nous avons défini une mesure de dissimilarité basée sur la structure des treillis et normalisée par leur largeur. Les treillis de Galois créent une hiérarchie sur la dualité extension / intention des concepts, qui est perdue lorsque les objets et les attributs sont considérés séparément, et notre mesure de dissimilarité ne prenait pas en compte cet aspect fondamental du treillis de concepts.

Le but de cet article est de présenter une nouvelle mesure de similarité entre treillis. Bien qu'il existe une abondante littérature portant sur des similarités entre graphes (orientés) (Ullmann, 1976), à notre connaissance il existe pas de littérature sur les similarités entre treillis de concepts (Domenach, 2015) prenant en compte leur dualité intrinsèque. En utilisant une

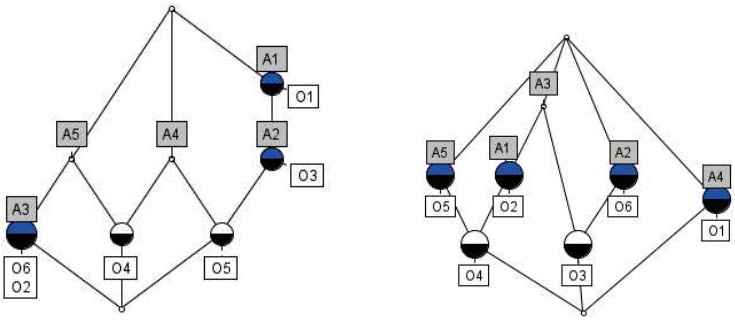

FIG. 1 – *Exemples de treillis* $\mathbb{L}_1$ *et* $\mathbb{L}_2$

approche de nature similaire à Champin et Solnon (2003), la méthode introduite ici est basée sur un modèle expressif utilisant des correspondances entre objets et entre attributs, correspondances qui peuvent ne pas être injectives ou surjectives. Chaque objet (attribut) du premier treillis peut être associé à un certain nombre d'objets (attributs) du second.

2 Analyse des Concepts Formels

Introduction Nous rappelons ici les notations standards de l'analyse des concepts formels (ACF) et nous renvoyons le lecteur à Ganter et Wille (1999) et à Caspard et al. (2012) pour des résultats sur les treillis comme ensembles ordonnés. Un *contexte* (G, M, I) est défini comme un ensemble G d'objets, un ensemble M d'attributs, et une relation binaire $I \subseteq G \times M$. $(G, m) \in I$ signifie que "l'objet g est liée avec l'attribut m par la relation I". Deux opérateurs de dérivation peuvent être définis sur les ensembles d'objets et d'attributs, $\forall O \subseteq G, A \subseteq M$, $O' = \{m \in M : \forall g \in O, (g, m) \in I\}$, $A' = \{g \in G : \forall m \in A, (g, m) \in I\}$. Ces deux opérateurs $(\cdot)'$ définissent une correspondance de Galois entre l'ensemble des parties de l'ensemble des objets $\mathcal{P}(G)$ et l'ensemble des parties des attributs $\mathcal{P}(M)$. Une paire $(O, A), O \subseteq G, A \subseteq M$, est un *concept formel* ssi $O' = A$ et $A' = O$. O est appelé *extension* et A *intention* du concept. La composition de ces deux opérateurs $(.)''$ forme un opérateur de fermeture sur $\mathcal{P}(G)$ (resp. $\mathcal{P}(M)$), créant un (double) isomorphisme entre les ensembles des fermés de $\mathcal{P}(G)$ et $\mathcal{P}(M)$.

Treillis de Galois L'ensemble des concepts formels est ordonné par inclusion des extensions (ou, dualement, par inclusion des intentions), *i.e.*, $(O_1, A_1) \leq (O_2, A_2)$ ssi $O_1 \subseteq O_2$ (ou dualement $A_2 \subseteq A_1$), et forme un treillis complet (Barbut et Monjardet, 1970), le *treillis de concepts* ou treillis de Galois noté $\mathbb{L} = \mathfrak{B}(G, M, I)$. Un diagramme de Hasse peut être associé au treillis de concepts comme le graphe de la relation de couverture : le concept (O_1, A_1) est couvert par (O_2, A_2), $(O_1, A_1) \prec (O_2, A_2)$, quand il n'existe pas de concept (O_3, A_3) tel que $(O_1, A_1) < (O_3, A_3) < (O_2, A_2)$. Dans le diagramme de Hasse, chaque concept du treillis est représenté sous la forme d'un sommet dans le plan et les arêtes vont vers le haut de (O_1, A_1) à

TAB. 1 – *Exemple de correspondance arbitraire c_o entre objets des treillis de la figure 1, chaque rangée étant un élément de G_1 et chaque colonne un élément de G_2*

c_o	o_1	o_2	o_3	o_4	o_5	o_6
o_1				X		
o_2				X		
o_3						
o_4	X			X		
o_5			X	X		
o_6				X		

(O_2, A_2) quand $(O_1, A_1) \prec (O_2, A_2)$. Toutes les figures de cet article ont été crées en utilisant le logiciel ConExp [1] (Yevtushenko, 2000).

3 Nouvelle Mesure de Similarité

Jaccard (1901) a créé une mesure de similarité simple, sur des ensembles, définie comme le ratio entre les éléments communs sur leur union, généralisée plus tard par Tversky (1977). Appliquée aux treillis de concepts, la mesure Jaccard peut être écrite comme :

$$sim(\mathbb{L}_1, \mathbb{L}_2) = \frac{f(descr(\mathbb{L}_1) \sqcap descr(\mathbb{L}_2))}{f(descr(\mathbb{L}_1) \sqcup descr(\mathbb{L}_2))} \; \forall \mathbb{L}_1, \mathbb{L}_2 \in \mathscr{L} \tag{1}$$

Cet article est centré au cas où f est la fonction de cardinalité, mais les résultats peuvent être facilement étendus à toute fonction positive et monotone non décroissante par rapport à l'ordre sur $\mathscr{L}$ ($\mathbb{L}_1 \sqsubseteq \mathbb{L}_2$ implique $f(\mathbb{L}_1) \leq f(\mathbb{L}_2)$). $descr$ est une fonction de description, qui peut être considérée comme un codage de treillis permettant une comparaison entre treillis. Dans les paragraphes suivants, nous définissons d'abord la description commune entre les deux treillis, notre numérateur, avant de spécifier leur union, notre dénominateur.

Correspondance Afin de définir une mesure de similarité entre deux treillis $\mathbb{L}_1 = \mathfrak{B}(G_1, M_1, I_1)$ et $\mathbb{L}_2 = \mathfrak{B}(G_2, M_2, I_2)$, nous définissons d'abord une correspondance (arbitraire) c_o (resp. c_a) qui lient les objets (resp. attributs) afin d'identifier leurs caractéristiques communes. Ces correspondances peuvent être considérées soit comme des connaissances spécialisées, correspondant à des caractéristiques d'un treillis à l'autre, ou comme un problème d'optimisation, cherchant la meilleure adéquation possible. Formellement, $c_o \in G_1 \times G_2$ est une relation binaire entre les objets de $\mathbb{L}_1$ et les objets de $\mathbb{L}_2$. Ce n'est pas une fonction entre G_1 et G_2 car tout objet peut avoir zéro, un ou plusieurs objets associés. Le tableau 1 montre un exemple de correspondance entre les objets des deux treillis de concepts de la figure 1.

Étant donné une correspondance c_o entre G_1 et G_2 associant zéro ou plusieurs objets de $\mathbb{L}_2$ avec chaque objet de $\mathbb{L}_1$, nous définissons l'image d'un objet $o_1 \in G_1$ comme $c_o(o_1) = \{o_2 \in G_2 : (o_1, o_2) \in c_o\}$. Cette définition peut être étendue à un ensemble d'objets $O_1 \subseteq G_1$ comme

1. Disponible à l'adresse `http://conexp.sourceforge.net/`

le produit cartésien des images de chacun des éléments de O_1 : $c_o(O_1) = \{\{y_1, y_2, ...\}, y_i \in c_o(o_i)\ \forall o_i \in O_1\}$ et $c_o(\emptyset) = \emptyset$. Par exemple, en utilisant la correspondance du tableau 1, $c_0(\{o_4, o_5\}) = \{\{o_1, o_3\}, \{o_1, o_4\}, \{o_3, o_4\}, \{o_4\}\}$. Des définitions similaires de l'image d'un attribut ou d'un ensemble d'attributs sont utilisés pour la correspondance c_a qui lie les attributs de M_1 et M_2.

Descriptions Communes Afin de définir la description commune entre deux treillis de concepts, nous devons d'abord définir de quelle façon l'information contenue dans $\mathbb{L}_1$ est représentée dans $\mathbb{L}_2$. Nous définissons cette information comme le ratio pour chaque concept de $\mathbb{L}_1$ d'être présent, au moins partiellement, dans $\mathbb{L}_2$ par la correspondance c_o.

Description sur les Objets Considérons le concept $\lambda = (O_1, A_1) \in \mathbb{L}_1$. La description des objets du concept λ de $\mathbb{L}_1$ à $\mathbb{L}_2$ selon la correspondance c_o, notée $descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2}(\lambda)$, est l'union de (O_1, a_1), $a_1 \in A_1$, de sorte que a_1 fasse partie d'un concept de $\mathbb{L}_2$ qui contient une image de O_1 par c_o. Cela correspond à l'information contenue dans le concept λ en fonction de son ensemble d'objets qui est présent, au moins partiellement, dans $\mathbb{L}_2$ par c_o. Formellement, $descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2}(\lambda) = \{(O_1, a_1), a_1 \in A_1,\ \exists X_1 \in c_o(O_1) : a_1 \in X_1'\}$. La description générale des objets de $\mathbb{L}_1$ dans $\mathbb{L}_2$ est l'union des descriptions de chaque concept :

$$descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2} = \bigcup_{\lambda \in \mathbb{L}_1} descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2}(\lambda)$$

La description commune entre deux treillis de concepts est alors l'union des descriptions d'un treillis à l'autre, à savoir $descr^{c_o}(\mathbb{L}_1) \sqcap descr^{c_o}(\mathbb{L}_2) = descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2} \cup descr^{c_o}_{\mathbb{L}_2 \to \mathbb{L}_1}$. Continuant avec notre exemple de la figure 1, en utilisant la correspondance c_o du tableau 1, on a $descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2}(\{o_3, o_5\}, \{a_1, a_2\}) = \{(\{o_3, o_5\}, \{a_1\}), (\{o_3, o_5\}, \{a_2\})\}$ vu que $c_o(\{o_3, o_5\}) = \{\{o_3\}, \{o_4\}\}$, $a_1 \in \{o_4\}'$ et $a_2 \in \{o_3\}'$. Mais, $descr^{c_o}_{\mathbb{L}_1 \to \mathbb{L}_2}(\{o_4, o_5\}, \{a_4\}) = \emptyset$ vu que $c_o(\{o_4, o_5\}) = \{\{o_1, o_3\}, \{o_1, o_4\}, \{o_3, o_4\}, \{o_4\}\}$ et le seul concept de $\mathbb{L}_2$ contenant a_4 est $(\{o_1\}, \{a_4\})$. Nous pouvons définir dualement description commune sur les attributs entre $\mathbb{L}_1$ et $\mathbb{L}_2$.

Description sur Objets et Attributs Aucunes des deux définitions précédentes de descriptions sur objets et sur attributs ne sont tout à fait satisfaisantes car elles considèrent les objets et les attributs séparément. Afin de tenir compte de la double nature des treillis de concepts, ces définitions de descriptions, soit sur des objets ou sur les attributs, conduisent à une description unifiée sur les deux dimensions. La description de $\mathbb{L}_1$ dans $\mathbb{L}_2$ sur les objets et attributs est une combinaison de la description des objets ainsi que la description des attributs, à savoir $\forall \lambda = (O_1, A_1) \in \mathbb{L}_1$:

$$\begin{aligned} descr^{c_o, c_a}_{\mathbb{L}_1 \to \mathbb{L}_2}(\lambda) = &\{(O_1, a_1),\ a_1 \in A_1,\ \exists X_1 \in c_o(O_1),\ \exists y_1 \in c_a(a_1) :\ y_1 \in X_1'\} \\ &\cup \{(o_1, A_1),\ o_1 \in O_1,\ \exists Y_1 \in c_a(A_1),\ \exists x_1 \in c_o(o_1),\ z_1 \in Y_1'\} \end{aligned} \quad (2)$$

De même, la description des objets et des attributs de $\mathbb{L}_1$ to $\mathbb{L}_2$ est définie comme : $descr^{c_o, c_a}_{\mathbb{L}_1 \to \mathbb{L}_2} = \bigcup_{\lambda \in \mathbb{L}_1} descr^{c_o, c_a}_{\mathbb{L}_1 \to \mathbb{L}_2}(\lambda)$. Continuant avec notre exemple de la figure 1, en utilisant la correspondance c_o du tableau 1 et c_a correspondance identité ($\forall i, c_a(a_i) = a_i', a_i \in$

$M_1, a'_i \in M_2$), on a $descr^{c_o,c_a}_{\mathbb{L}_1 \to \mathbb{L}_2}(\{o_3, o_4\}, \{a_1, a_5\}) = \{(\{o_3, o_4\}, \{a_1\}), (\{o_3, o_4\}, \{a_2\}), (\{o_4\}, \{a_1, a_2\})\}$.

La description commune de $\mathbb{L}_1$ et $\mathbb{L}_2$, utilisée comme le numérateur dans l'équation 1, est l'union de la description de $\mathbb{L}_1$ dans $\mathbb{L}_2$ et de la description de $\mathbb{L}_2$ dans $\mathbb{L}_1$. Cet ensemble contient toutes les caractéristiques présentes à la fois dans $\mathbb{L}_1$ et $\mathbb{L}_2$ qui sont partiellement identifiées par les correspondances c_o et c_a. La description commune des deux réseaux$\mathbb{L}_1$ et $\mathbb{L}_2$ est définie comme :

$$descr(\mathbb{L}_1) \sqcap descr(\mathbb{L}_2) = descr^{c_o,c_a}_{\mathbb{L}_1 \to \mathbb{L}_2} \cup descr^{c_o,c_a}_{\mathbb{L}_2 \to \mathbb{L}_1} \tag{3}$$

Union des Descriptions Afin de compléter notre définition de la similarité de Jaccard entre deux treillis, nous avons besoin de définir l'union des descriptions de ces treillis comme :

$$\begin{aligned} descr(\mathbb{L}_1) \sqcup descr(\mathbb{L}_2) = & \bigcup_{(O_1,A_1)\in\mathbb{L}_1} \{ \bigcup_{a_1\in A_1} \{(O_1, a_1)\} \cup \bigcup_{(o_1\in O_1)} \{(o_1, A_1)\} \} \\ & \cup \bigcup_{(O_2,A_2)\in\mathbb{L}_2} \{ \bigcup_{a_2\in A_2} \{(O_2, a_2)\} \cup \bigcup_{(o_2\in O_2)} \{(o_2, A_2)\} \} \end{aligned} \tag{4}$$

Scissions Un problème de cette approche basée sur des correspondances est que c_o et c_a sont des relations binaires, pas des fonctions. Ainsi, tout objet (attribut) de $\mathbb{L}_1$ ou $\mathbb{L}_2$ peut avoir plus d'une image. Prenons le cas extrême où chaque objet / attribut de $\mathbb{L}_1$ est lié à tout autre objet / attribut de $\mathbb{L}_2$. Bien que très peu instructive, la similarité, telle que définie dans l'équation 1, sera artificiellement élevée. Les scissions sont définies quand un objet ou un attribut a plus d'une image en c_o ou c_a. Informellement, les scissions quantifient le manque de précision dans c_o et c_a. Plus un objet (resp. un attribut) a d'images par c_o (resp. c_a), moins il est informatif. Nous pouvons maintenant revenir à notre mesure de similarité de l'équation 1 en prenant les scissions en compte, vu que nous voulons avoir une mesure de similarité qui sera diminuée à mesure que le nombre de scissions augmente. La mesure de similarité est définie comme suit :

$$sim_{c_o,c_a}(\mathbb{L}_1, \mathbb{L}_2) = \frac{f(descr(\mathbb{L}_1) \sqcap descr(\mathbb{L}_2)) - g(splits(c_o) \cup splits(c_a))}{f(descr(\mathbb{L}_1) \sqcup descr(\mathbb{L}_2))} \tag{5}$$

avec f et g deux fonctions positives et non décroissante (ici des cardinalités). C'est une mesure de similarité car nous avons $sim_{c_o,c_a}(\mathbb{L}_1, \mathbb{L}_1) = 1$ avec diagonal c_a, c_o. Cependant, elle n'est pas normalisée en raison de scissions possibles, et ainsi peut devenir négatif. Bien que différentes définitions des scissions peuvent être utilisées pour cette mesure de similarité, nous nous sommes concentrés sur le cas particuliers où $split$ est le nombre d'images moins un dans c_o et c_a, i.e. $\Sigma_{x\in G_1\cup G_2}(|c_o(x)| - 1) + \Sigma_{y\in M_1\cup M_2}(|c_a(y)| - 1)$, et nous renvoyons le lecteur à Domenach et Rajabi (2015) pour une discussion sur différentes mesures de scissions.

4 Conclusion et Perspectives

Dans cet article, nous avons présenté une adaptation de la mesure de similarité de Champin et Solnon (2003) pour graphes orientés dans le cadre de l'ACF. Basée sur mesure de Jaccard, notre similarité utilise les correspondances entre objets et entre attributs des deux treillis de

concepts. Elle est en mesure de saisir le rôle liés entre intention et extension, entre objets et attributs entre les deux treillis de concepts considérés. Vu que notre mesure de similarité repose sur la recherche des meilleures correspondances c_o et c_a, une question clé est l'existence d'algorithmes efficaces pour leur mise en oeuvre. Comme f et g sont croissantes, il est difficile d'évaluer le changement de similarité lorsque une des correspondances augmente. Une investigation future est l'évaluation statistique de la mesure de similarité en fonction de f et g. Nous prévoyons également d'analyser son comportement et de la corréler avec des mesures de similarité existantes.

Références

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

Caspard, N., B. Leclerc, et B. Monjardet (2012). *Finite ordered sets : concepts, results and uses*. Number 144. Cambridge University Press.

Champin, P.-A. et C. Solnon (2003). Measuring the similarity of labeled graphs. In K. Ashley et D. Bridge (Eds.), *Case-Based Reasoning Research and Development*, pp. 80–95. Springer Berlin Heidelberg.

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 Z. Rajabi (2015). Correspondence-based lattice similarity measure. In *Proceedings of European Conference on Data Analysis 2015*.

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.

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

Ullmann, J. R. (1976). An algorithm for subgraph isomorphism. *Journal of the ACM 23*(1), 31–42.

Yevtushenko, S. A. (2000). System of data analysis "concept explorer". (in russian). In *Proceedings of the 7th national conference on Artificial Intelligence KII-2000*, pp. 127–134.

Summary

This paper is in the formal concept analysis framework, an algebraic hierarchisation method of data based on the notion of extent / intent, i.e. of maximally shared attributes and objects. Here we present a correspondence-based similarity measure between two formal concept lattices, defined on an expressive model using correspondences between objects and between attributes of the two lattices. A key point of our approach is that the correspondences may not be mappings and may associate each object (resp. attribute) of one lattice with several objects (resp. attributes) of another one.