Connaissances de domaine et treillis de concepts pour l'exploration progressive de données complexes

Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, Malika Smaïl-Tabbone

Connaissances de domaine et treillis de concepts pour l'exploration progressive de données complexes

Nizar Messai[1 †], Marie-Dominique Devignes[2], Amedeo Napoli[2], and Malika Smaïl-Tabbone[2]

[1] Procton Labs, 180 rue de Vaugirard 75015 Paris
nizar.messai@proctonlabs.org
[2] LORIA UMR 7503, BP 239, 54506 Vandœuvre-Lès-Nancy
{devignes, napoli, malika}@loria.fr

Résumé :

Nous présentons dans cet article l'Analyse de Concepts Formels par Similarité, qui adapte et étend l'Analyse de Concepts Formels classique à des données complexes, en s'appuyant sur des connaissances de domaine. Ces connaissances sont considérées pour définir la similarité entre les données qui se présentent sous la forme d'un contexte multivalué. En s'appuyant sur la similarité définie, les données du contexte sont groupées dans des concepts multivalués qui forment des treillis de concepts multivalués. La variation des critères dans la définition de la similarité aboutit à la modification de la structure de treillis obtenue et du niveau de précision dans les concepts. Nous exploitons cet aspect pour définir une méthode d'exploration progressive de données complexes par treillis de concepts multivalués. Nous détaillons l'application de cette méthode à l'organisation et à l'identification des sources de données biologiques de l'annuaire BioRegsity.

Mots-clés : Analyse de Concepts Formels, Connaissances de domaine, données complexes, ressources biologiques.

1 Introduction

Les développements et les travaux effectués dans le domaine du Web sémantique ont abouti à la mise en place de plusieurs ressources sémantiques fiables telles que la *Gene Ontology*, les thésaurus *MeSH* [1] et *WordNet*, la taxonomie *NCBI* [2], etc. Parallèlement à cela, les sources de données disponibles sur le Web ont continué leur évolution avec des contenus de plus en plus complexes, volumineux et hétérogènes. Ces évolutions ne sont

†. Nizar Messai a participé au travail de recherche présenté dans cet article en tant que doctorant de l'équipe LORIA - Orpailleur et ATER à l'UHP - Nancy 1 avant Août 2009

1. http ://www.nlm.nih.gov/mesh/meshhome.html
2. http ://www.ncbi.nlm.nih.gov/

pas totalement indépendantes. En effet, d'un coté la mise en place de sources de données s'appuie plus que jamais sur les ressources sémantiques, depuis la conception jusqu'à la structuration ou l'annotation du contenu, et de l'autre côté, les ressources sémantiques s'enrichissent au fur et à mesure que le contenu des sources de données s'enrichit et ce en définissant de nouveaux concepts liés à des nouvelles données devenues d'actualité dans les sources de données. Le défi aujourd'hui est de s'appuyer sur ce couplage pour exploiter de la meilleure façon les données disponibles sur le Web. Dans cet article, nous abordons cette problématique dans le but d'assurer une meilleur exploitation des données de l'annuaire BioRegistry (Devignes *et al.*, 2008).

BioRegistry référence plus de mille sources de données biologiques disponibles sur le Web. Ces sources de données sont annotées par des métadonnées définies en adéquation avec le modèle de référence DCMI [3] et prises à partir de ressources sémantiques telles que le thésaurus MeSH et la taxonomie NCBI. Dans un travail antérieur (Messai *et al.*, 2008a), nous avons défini une méthode qui s'appuie sur l'Analyse de Concepts Formels (ACF) (Ganter & Wille, 1999) pour l'exploitation du contenu de BioRegistry tout en considérant les relations sémantiques entre les données de cet annuaire. L'application de l'ACF a nécessité la transformation des données de BioRegistry sous la forme d'un contexte binaire. Cette transformation entraîne souvent la perte des relations sémantiques entre les données d'origine. Nous avons alors tenté de faire appel à ces relations, à posteriori à l'étape d'application de l'ACF, sous la forme de préférences dans des requêtes et sous la forme de dépendances entre attributs.

Dans cet article nous proposons une nouvelle approche, l'ACF par similarité, qui prend en compte les relations sémantiques au moment même de l'application de l'ACF. Il s'agit d'adapter l'ACF à des données sous la forme d'un contexte multivalué, en considérant les relations sémantiques entre ces données. La suite de l'article est organisée comme suit. La section 2 rappelle les définitions de base de l'ACF. La section 3 donne les détails de la formalisation de l'ACF par similarité. La section 4 montre l'apport de l'ACF par similarité dans l'exploration de données complexes et donne un aperçu sur l'implémentation de l'approche et les premiers résultats de l'experimentation sur des données réelles de BioRegistry. Finalement la section 5 conclut l'article en faisant un bilan des travaux et en présentant les perspectives ouvertes par cette approche.

2 Analyse de Concepts Formels et données complexes

2.1 Analyse de Concepts Formels (ACF)

L'ACF s'applique sur des données qui se présentent sous la forme d'un *contexte formel* : un triplet $\mathbb{K} = (G, M, I)$ où G est un ensemble d'objets, M est un ensemble d'attributs et I est une relation binaire entre G et M ($I \subseteq G \times M$). Un couple $(g, m) \in I$ (notée aussi par gIm) signifie que l'objet $g \in G$ possède l'attribut $m \in M$. La table 1 donne un exemple de contexte formel. Les objets sont des sources de données biologiques et les attributs sont de trois types : le sujet du contenu de chaque source de données (des termes provenant du MeSH), l'organisme concerné (des termes prove-

3. http ://dublincore.org/

nant du NCBI) et la catégorie de la source de données (des numéros de catégories ou sous-catégories du *Nucleic Acid Research*[4] (NAR). La relation I exprime le fait qu'une source de données est annotée par un attribut (auquel cas la case correspondante dans le tableau contient "$\times$") ou non (auquel cas la case correspondante est vide).

Table 1 – Un exemple de contexte formel.

	Sujet (MeSH)				Organisme d'intérêt (NCBI)							Catégorie (NAR)			
	Genome	Genome Components	Proteins	Transcription factors	Eukaryotes	Human	Mammals	Plants	Prokaryotes	Rice	Vertebrates	1.2	5.2	7.3	13
ExInt		×			×							×			
HSD			×			×								×	
rRNDB	×								×				×		
SpliceDB		×					×					×			
CropNet								×							×
GOLD	×												×		
INE										×					×
TRANSCompel				×							×	×			

A partir d'un contexte formel, l'ACF permet de grouper les objets en fonction des attributs qu'ils ont en commun et, de façon duale, les attributs en fonction des objets qui les possèdent. Ce groupement est effectué comme suit. Pour chaque ensemble d'objets $A \subseteq G$, A' = $\{m \in M \mid \forall g \in A, gIm\}$ est l'ensemble d'attributs communs à tous les objets de A. De façon duale, pour chaque ensemble d'attributs $B \subseteq M$, B' = $\{g \in G \mid \forall m \in B, gIm\}$ est l'ensemble d'objets possédant tous les attributs de B.

Les groupements d'objets et d'attributs donnent lieux à des *concepts formels*. Un concept formel est un couple (A, B) tel que $A \subseteq G$, $B \subseteq M$, $A' = B$ et $B' = A$. A et B sont respectivement appelés $extension$ et $intension$ du concept formel (A, B). L'ensemble des concepts formels associés au contexte formel $\mathbb{K} = (G, M, I)$ est noté par $\mathfrak{B}(G, M, I)$. Les concepts de $\mathfrak{B}(G, M, I)$ sont ordonnés par une relation de subsomption entre concepts (notée par $\sqsubseteq$) qui se définit par : $(A1, B1) \sqsubseteq (A2, B2)$ si et seulement si $A1 \subseteq A2$ (ou de façon duale $B2 \subseteq B1$), $(A1, B1)$ et $(A2, B2)$ étant deux concepts formels de $\mathfrak{B}(G, M, I)$. $(A2, B2)$ est dit *subsumant* ou *super-concept* de $(A1, B1)$ et $(A1, B1)$ est dit *subsumé* ou *sous-concept* de $(A2, B2)$. Cette relation de subsomption permet d'organiser les concepts formels en un treillis complet $(\mathfrak{B}(G, M, I), \sqsubseteq)$ appelé *treillis de concepts* ou encore treillis de Galois et noté par $\underline{\mathfrak{B}}(G, M, I)$. Le treillis de concepts correspondant au contexte formel donné dans la table 1 est donné dans la figure 1.

4. http ://www.oxfordjournals.org/nar/database/c

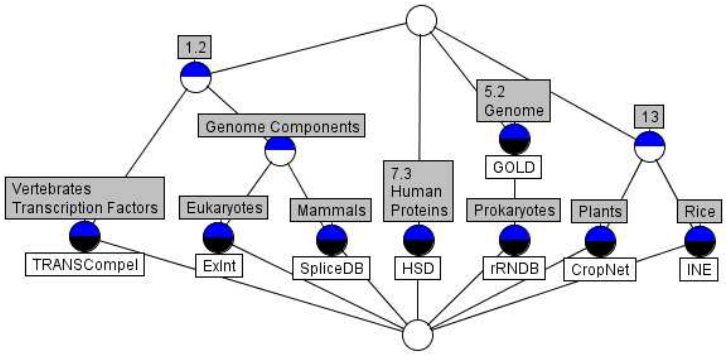

Fig. 1 – Le treillis de concepts correspondant au contexte formel donné dans la table 1.

2.2 ACF et données complexes : les contextes multivalués

Les approches qui s'appuient sur l'ACF sont définies pour l'analyse et l'exploitation des données du monde réel. Ces données ne se présentent pas forcément sous la forme de contextes binaires. Ceci est le cas des données de BioRegistry que nous considérons dans ce travail. Les sources de données sont annotées par des métadonnées réparties sur un ensemble d'attributs tels que le sujet de la base, l'organisme étudié, la catégorie de la base, etc. La représentation tabulaire directe de ces données produit un tableau où les lignes sont les objets, les colonnes sont les attributs et où chaque case contient la valeur que l'attribut en colonne prend pour l'objet en ligne. En ACF, cette représentation est appelée *contexte multivalué* (Ganter & Wille, 1999) et est définie comme étant un quadruplet (G, M, W, I) où G est un ensemble d'objet, M un ensemble d'attributs, W un ensemble de valeurs d'attributs et I est une relation ternaire entre G, M et W (i.e., $I \subseteq G \times M \times W$). $(g, m, w) \in I$ (noté aussi par $m(g) = w$) signifie que "*l'attribut* m *prend la valeur* w *pour l'objet* g". La table 2 donne un exemple de contexte multivalué.

Table 2 – Un exemple de contexte multivalué.

Attributs / Objets	Sujet (MeSH)	Organisme d'intérêt (NCBI)	Catégorie (NAR)
ExInt	Genome components	Eukaryotes	1.2
HSD	Proteins	Human	7.3
rRNDB	Genome	Prokaryotes	5.2
SpliceDB	Genome components	Mammals	1.2
CropNet		Plants	13
GOLD	Genome		5.2
INE		Rice	13
TRANSCompel	Transcription factors	Vertebrates	1.2

2.3 Approches existantes pour le traitement de contextes multivalués

L'application de l'ACF à des contextes multivalués nécessite d'abord de les transformer en contextes formels. Cette transformation est appelée *échelonnage conceptuel* (Ganter & Wille, 1999) et consiste à transformer chaque attribut multivalué en un ensemble d'attributs binaires. De cette manière, l'exemple de contexte formel donné dans la table 1 est le résultat d'un certain échelonnage conceptuel du contexte multivalué donné dans la table 2. Dans cet exemple, l'échelonnage conceptuel, appelé échelonnage plan (*plain scaling*) (Ganter & Wille, 1999), consiste à transformer chacune des valeurs distinctes des attributs multivalués en un attribut binaire du contexte formel résultat. En dehors de certains types particuliers d'échelonnage et des règles permettant des les obtenir, la définition d'un échelonnage pour un contexte multivalué n'est pas unique. Le choix d'un échelonnage parmi d'autres reste très subjectif et dépend de l'interprétation des attributs des contextes multivalués. De cet fait, cette transformation des contextes multivalués est souvent difficile à automatiser lorsque ces contextes sont volumineux et représentent des données complexes et variées. De plus, la répartition des valeurs d'un même attribut multivalué en un ensemble d'attributs binaires indépendants est un facteur qui introduit des biais dans les données à manipuler avant même de pouvoir appliquer l'ACF.

Dans la littérature, certaines approches ont été proposées pour étendre l'ACF aux contextes multivalués sans avoir à procéder à l'échelonnage conceptuel. Ces approches consistent à redéfinir les opérateurs de dérivations entre les objets et les attributs en adéquation avec les contextes multivalués étudiés et aboutissent ainsi à la définition de nouvelles connections de Galois. Dans (Brito & Polaillon, 2005), des définitions et résultats relatifs au données symboliques sont utilisés pour étudier les contextes multivalués représentant des données intervalles, histogrammes et probabilités. Dans (Belohlavek & Vychodil, 2005), un ensemble d'approches similaires, utilisant les résultats et définitions de la logique floue, sont présentées et comparées. Elles sont définies pour les contextes formels flous (contextes multivalués où les valeurs sont des degrés de vérité de la relation entre les objets et les attributs).

Ces approches, bien qu'elles permettent d'étendre les résultats de l'ACF à des contextes multivalués, restent applicables à des données sous des formats bien particuliers. Contrairement à ces approches, l'Analyse de Concepts Logiques (Ferré & Ridoux, 2000) et les *Pattern Structures* (Ganter & Kuznetsov, 2001) présentent de vraies généralisations de l'ACF aux données complexes. Ces deux approches s'appuient sur la description des objets par l'intermédiaire de descriptions formelles généralisant les attributs multivalués, ce qui leur permet d'être génériques. Cependant, leur application à des données réelles nécessite de les instancier en adéquation avec chaque type de données.

Dans (Messai *et al.*, 2008b), une nouvelle approche plus intuitive a été proposée pour étendre l'ACF aux contextes multivalués où les valeurs d'attributs appartiennent à des domaines munis d'une relation d'ordre total, comme c'est le cas pour les données numériques. Cette approche s'appuie sur la similarité entre les valeurs de chacun des attributs d'un contexte multivalué pour effectuer les groupements d'objets. Dans la suite de cet article nous étendons cette approche aux contextes multivalués où les domaines de valeurs ne sont pas forcément munis d'un ordre total. Nous nous appuyons sur des connaissances de domaine pour calculer la similarité entre les valeurs des attributs des

contextes et grouper, par la suite, les objets ayant les valeurs d'attributs les plus similaires.

3 Analyse de Concepts Formels par similarité

3.1 La similarité dans un contexte multivalué

L'intuition de base de l'ACF par similarité est la suivante : deux objets ou plus sont regroupés autour d'un attribut si et seulement si les valeurs de cet attribut pour ces objets sont similaires. La notion de similarité ici est générique et peut être définie selon les données étudiées et selon le domaine d'application. Dans le cadre de ce travail nous nous appuyons sur les connaissances de domaine qui se présentent sous la forme de hiérarchies de termes (MeSH, NCBI et NAR) pour calculer la similarité entre les valeurs des attributs. Les fragments de ces hiérarchies correspondant au contexte multivalué donné dans la table 2 sont donnés dans la figure 2.

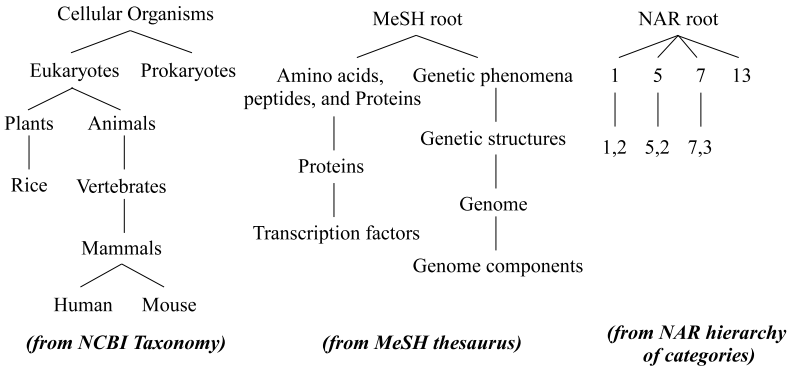

Fig. 2 – Fragments de hiérarchies contenant les valeurs des attributs du contexte multivalué donné dans la table 2.

Il existe plusieurs méthodes de calcul de la similarité dans une hiérarchie de termes (ou de manière générale hiérarchie de concepts, ontologie). Des comparaisons détaillées des principales méthodes ont été faites dans (Ganesan *et al.*, 2003) et (Hliaoutakis *et al.*, 2006). Dans la suite, nous nous appuyant sur le Coefficient de Jaccard et sur les hiérarchies de termes pour calculer la similarité entre les valeurs d'attributs du contexte. Pour cela, chacune des hiérarchies de termes est notée comme suit : $\mathcal{H} = (\mathcal{V}, E)$ où $\mathcal{V}$ est l'ensemble de termes, E est l'ensemble de liens entre ces termes et pour tout $v \in \mathcal{V}$, $Ancestors(v)$ est l'ensemble des termes sur le chemin allant de v à la racine de $\mathcal{H}$. La similarité entre deux termes v_i et $v_j \in \mathcal{V}$ est obtenue comme suit :

$$sim(v_i, v_j) = \frac{|Ancestors(v_i) \cap Ancestors(v_j)|}{|Ancestors(v_i) \cup Ancestors(v_j)|}$$

Deux termes v_i et v_j d'une même hiérarchie sont dit similaires, et on écrit $v_i \simeq v_j$, si et seulement si leur similarité est supérieure à un seuil $\theta \in [0, 1]$ donné, $sim(v_i, v_j) \geq \theta$. Lorsque $sim(v_i, v_j) < \theta$, v_i et v_j sont dits non similaires et on écrit $v_i \not\simeq v_j$.

3.2 Partage d'attributs entre objets

En s'appuyant sur cette définition de similarité entre les valeurs d'attributs, on définit le partage d'attributs entre objets dans un contexte multivalué (G, M, W, I) comme suit. Deux objets $g_i, g_j \in G$ partagent un attribut $m \in M$ si et seulement si $m(g_i) = w_i \neq \emptyset$, $m(g_j) = w_j \neq \emptyset$ et $m(g_i) \simeq m(g_j)$. On dit que les objets g_i et g_j partagent m pour les valeurs $\{w_i, w_j\}$ et on écrit $(m, \{w_i, w_j\})$. De manière générale, un ensemble d'objet $A \subseteq G$ partage (m, W_m^A), où $W_m^A = \{m(g) \in W,\ g \in A\}$ si et seulement si $\forall g_i, g_j \in A,\ m(g_i) \simeq m(g_j)$. Dans ce cas, on dit que A est valide pour m et que W_m^A est un ensemble de valeurs similaires de m pour A.

Par exemple, pour $\theta = 0.5$, les objets *HSD* et *SpliceDB* partagent (*NCBI*, {*Human, Mammals*}) puisque *Human*$\simeq$*Mammals* dans la taxonomie NCBI. Par contre, ces deux objets ne partagent pas l'attribut *MeSH* puisque *Proteins*$\not\simeq$*Genome components*.

Étant donné un contexte multivalué (G,M,W,I), un objet $g_j \in G$ est dit atteignable à partir d'un autre objet $g_i \in G$ pour un attribut $m \in M$ lorsque $m(g_i) \simeq m(g_j)$. De manière générale, l'ensemble d'objets atteignables à partir de $A \subseteq G$ est défini comme suit :

$$\mathfrak{R}(A, m) = \{g_i \in G \mid m(g_i) \simeq m(g),\ \forall g \in A\}$$

L'ensemble d'objets atteignables à partir de A pour un ensemble d'attributs $B \subseteq M$ est :

$$\mathfrak{R}(A, B) = \bigcap_{m \in B} \mathfrak{R}(A, m)$$

L'ensemble $\mathfrak{R}(A, m)$ peut ne pas être valide pour m du fait de la non transitivité de l'opérateur de similarité. Nous nous intéressons pour cela à l'ensemble valide maximal contenant A dans $\mathfrak{R}(A, m)$. Cet ensemble est obtenu en enlevant de $\mathfrak{R}(A, m)$ toute paire d'objets g_i, g_j telle que $m(g_i) \not\simeq m(g_j)$) :

$$\mathfrak{R}_v(A, m) = \mathfrak{R}(A, m) \setminus \{g_i \in \mathfrak{R}(A, m) \mid \exists g_j \in \mathfrak{R}(A, m)\ et\ m(g_i) \not\simeq m(g_j)\}$$

Plus généralement, l'ensemble valide maximal contenant $A \subseteq G$ pour $B \subseteq M$ est :

$$\mathfrak{R}_v(A, B) = \bigcap_{m \in B} \mathfrak{R}_v(A, m)$$

Lorsque un ensemble d'objets $A \subseteq G$ partage un attribut $m \in M$ (i.e. $\mathfrak{R}(A, m) \neq \emptyset$), l'ensemble maximal de valeurs similaires de m pour A est

$$\gamma(A, m) = \{m(g) \in W,\ g_i \in \mathfrak{R}_v(A, m)\}$$

et on dit que les objets dans A partagent l'attribut multivalué $(m, \gamma(A, m))$. Par exemple, pour $\theta = 0.5$, {*ExInt, CropNet, INE*} partage (*NCBI*, {*Eukaryotes, Plants, Rice*}).

Les couples de la forme (m, V_m) désignant un attribut $m \in M$ et un ensemble $V_m \subseteq W$ de valeurs possibles de m sont des éléments de $M \times \mathfrak{P}(W)$, $\mathfrak{P}(W)$ étant l'ensemble de parties de W. Ces éléments peuvent être partiellement ordonnés comme suit :

$$\{(m, V_1)\} \subseteq_\theta \{(m, V_2)\} \text{ si et seulement si } V_2 \subseteq V_1$$

et de manière générale pour $B_1 = \{(m_1^i, V_1^i)\}_{i \in I}$ et $B_2 = \{(m_2^j, V_2^j)\}_{j \in J}$ inclus dans $M \times \mathfrak{P}(W)$,

$B_1 \subseteq_\theta B_2$ si et seulement si $\forall (m_1^i, V_1^i) \in B_1, \exists (m_2^j, V_2^j) \in B_2 | m_1^i = m_2^j$ et $V_2^j \subseteq V_1^j$

Par exemple, {(*NCBI*,{*Human, Mammal, Vertebrates*})} $\subseteq_\theta$ {(*NCBI*,{Mammals, Vertebrates}), (*NAR*,{1.2})} pour $\theta = 0.5$.

L'opérateur "$\subseteq_\theta$" définit un ordre partiel sur les parties de $M \times \mathfrak{P}(W)$ puisqu'il résulte d'un produit de deux opérateurs d'ordre partiel qui sont l'inclusion sur M et l'inclusion sur W.

3.3 Opérateurs de dérivation et connexion de Galois

Étant donné un contexte multivalué (G, M, W, I), l'ensemble d'attributs multivalués communs à un ensemble d'objets $A \subseteq G$ est :

$$A^\uparrow = \{(m, \gamma(A, m)) \in M \times \mathfrak{P}(W) \mid \gamma(A, m) \neq \emptyset\}$$

De façon duale, l'ensemble d'objets partageant les attributs valués dans $B \subseteq M \times \mathfrak{P}(W)$ est :

$$B^\downarrow = \mathfrak{R}_v(\{g \in G \mid \forall\ (m, V_m) \in B,\ m(g) \simeq w,\ \forall w \in V_m\},\ B)$$

Par exemple, pour $\theta = 0.5$, {*INE*,*CropNet*}$^\uparrow$ = {(*NCBI*,{*Plants,Rice*}), (*NAR*,{*13*})} et {(*NCBI*,{*Plants,Rice*}), (*NAR*,{*13*})}$^\downarrow$ = {*INE*,*CropNet*}.

Les deux opérateurs $^\uparrow$ et $^\downarrow$ forment une connexion de Galois entre $(\mathfrak{P}(G), \subseteq)$ et $(\mathfrak{P}(M \times \mathfrak{P}(W)), \subseteq_\theta)$ appelée connexion de Galois par similarité.

3.4 Concepts multivalués et treillis de concepts multivalués

De façon analogue à l'ACF, nous définissons un concept multivalué comme étant un couple (A, B) où $A \subseteq G$ et $B \subseteq M \times \mathfrak{P}(W)$ tels que $A^\uparrow = B$ et $B^\downarrow = A$. A et B sont respectivement appelés extension et intension de (A, B). Par exemple, pour $\theta = 0.5$, ({*INE*,*CropNet*}, {(*NCBI*,{*Plants,Rice*}),(*NAR*,{*13*})}) est un exemple de concept multivalué.

Un concept multivalué (A_1, B_1) est dit sous-concept de (A_2, B_2) lorsque $A_1 \subseteq A_2$ (ce qui est équivalent à $B_2 \subseteq_\theta B_1$). Dans ce cas on dit que (A_2, B_2) est un super-concept de (A_1, B_1) et on écrit $(A_1, B_1) \leq_\theta (A_2, B_2)$. L'ensemble de tous les concepts multivalués d'un contexte multivalué partiellement ordonnés par l'intermédiaire de "$\leq_\theta$" forme un treillis de concepts multivalués pour le seuil de similarité θ. Ce treillis est noté par $\mathfrak{B}_\theta(G, M, W, I)$.

Le treillis de concepts multivalués correspondant au contexte multivalué donné dans la table 2 pour un seuil de similarité $\theta = 0.5$ est donné dans la figure 3.

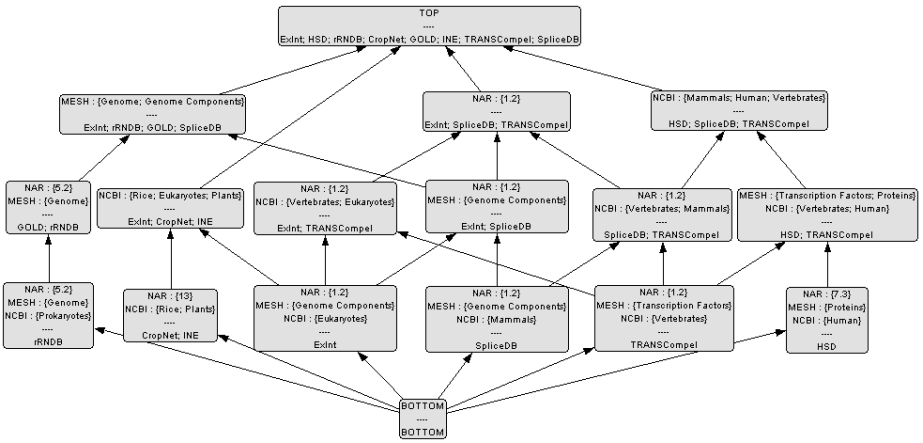

Fig. 3 – Le treillis de concepts multivalué $\underline{\mathfrak{B}}_{0.5}(G, M, W, I)$.

4 Application à l'exploration de données complexes

4.1 Exploration progressive des contextes multivalués

Le treillis de concepts multivalués donné dans la figure 3 est obtenu pour un seuil de similarité particulier ($\theta = 0.5$). La variation de ce seuil a une influence directe sur le partage d'attributs entre les objets dans un contexte multivalué. La diminution de θ constitue une relaxation de la condition du partage d'attributs. De plus en plus d'objets peuvent ainsi être groupés autour d'attributs à valeurs peu similaires. Dans ce cas on dit que les concepts multivalués obtenus sont de granularité large ou grossière. La granularité la plus large est obtenue pour $\theta = 0$ qui signifie que toutes les valeurs d'un même attribut sont similaires.

Dans le cas contraire, l'augmentation de θ constitue un raffermissement de la condition du partage d'attributs. Seuls peu d'objets peuvent être groupés autour d'attributs à valeurs identiques ou très similaires. Dans ce cas les concepts multivalués obtenus sont de granularité fine. La granularité la plus fine est obtenue pour $\theta = 1$ auquel cas seuls les objets ayant des valeurs identiques d'un attribut peuvent être groupés autour de cet attribut.

Cette possibilité de faire varier la granularité des concepts et leur nombre dans les treillis de concepts multivalués est particulièrement intéressante dans le cas de l'exploration de contextes multivalués volumineux. L'exploration peut commencer par la construction d'un treillis peu précis en choisissant un faible seuil de similarité. Ce premier treillis donne une vision globale sur les données dans le contexte. En suite, selon le besoin en précision, on choisi des seuils de plus en plus élevés pour générer des treillis de plus en plus détaillés avec des concepts de granularité de plus en plus fine.

4.2 Zooms et navigation dynamique

Lorsqu'un seuil convenable est choisi, le treillis de concepts multivalués correspondant peut servir de support pour la navigation en suivant la hiérarchie de ses concepts de la même manière que dans le cas de l'ACF classique (Carpineto & Romano, 1996). L'avantage principal apporté par les treillis de concepts multivalués dans ce cadre est la possibilité d'effectuer des zooms dynamiques sur certaines parties du treillis pendant la navigation. En effet, pendant la navigation, on peut être intéressé par un concept particulier qu'on voudrait voir en détail sans avoir à reconstruire tout le treillis et à recommencer la navigation. Dans ce cas, l'opération de zoom-avant permet de basculer dynamiquement vers la partie correspondante à ce concept dans un treillis correspondant à un seuil plus élevé. Dans un autre cas de figure, on pourrait être intéressé par un zoom-arrière auquel cas on considère un treillis correspondant à un seuil plus petit. Ce passage d'un treillis à l'autre nécessite que différents treillis correspondants à différents seuils doivent être générés préalablement. Cette solution peut être coûteuse en ressources mémoire lorsque les treillis sont de grande taille. Une autre manière de procéder consiste alors à extraire la partie du contexte correspondante à la partie du treillis en cours de visualisation et à construire le treillis correspondant pour un seuil plus petit ou plus grand selon le type de zoom demandé. L'illustration des deux types de zoom est donnée dans la figure 4. Les treillis considérés sont $\underline{\mathfrak{B}}_{0.5}(G, M, W, I)$ (gauche) et $\underline{\mathfrak{B}}_{0.2}(G, M, W, I)$ (droite). Le zoom-avant consiste à passer du treillis à droite au treillis à gauche et le zoom-arrière est donné par le chemin inverse.

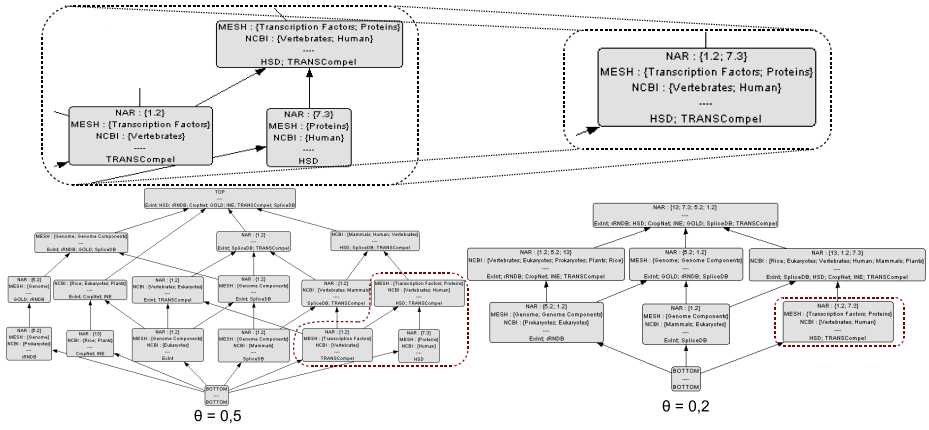

Fig. 4 – Un exemple de zoom dans un processus de navigation dynamique.

4.3 Experimentation

L'ACF par similarité introduite dans cet article est implémentée dans le prototype *SimBA* [5]. La construction du treillis s'appuie sur une extension de la méthode de Malgrange pour le calcul des concepts par intersections successives des lignes du contexte.

L'expérimentation de SimBA a été faite sur un corpus extrait de BioRegistry. Nous avons considéré dans un premier temps un contexte multivalué formé par 48 sources de données, décrites par 5 attributs multivalués relatifs aux sujets de ces sources et correspondant à 5 différentes branches du MeSH et 39 termes MeSH distincts répartis dans le contexte avec une densité de 0.34. Ensuite, l'expérimentation a été étendue à toutes les sources de données décrites par des termes MeSH dans BioRegistry. Le contexte multivalué correspondant est constitué de 585 sources de données, 9 attributs correspondant à 9 différentes hiérarchies du MeSH et 410 termes MeSH distincts repartis dans le contexte avec une densité de 0.3.

Dans les deux cas, nous avons calculé la similarité entre les termes en s'appuyant sur la hiérarchie du MeSH et en appliquant la formule introduite dans la section 3.1. Ce calcul est facilité par les "*Tree Numbers*" associés aux termes MeSH. Par exemple, pour les termes *Mammals* et *Vertebrates* ayant respectivement "*B01.150.900.649*" et "*B01.150.900*" comme *Tree Number*, on peut directement constater qu'ils ont 3 ancêtres communs (*B01*, *B01.150* et *B01.150.900*) et que leur similarité est de 0.75.

L'évolution du nombre de concepts dans les treillis obtenus en fonction du seuil de similarité est donnée dans la figure 5. On remarque sur ces graphiques que l'allure est la même pour les trois exemples considérés malgré leur importante différence de taille. Ceci confirme l'applicabilité des idées présentées dans cet article à des données complexes dans le cadre d'applications réelles.

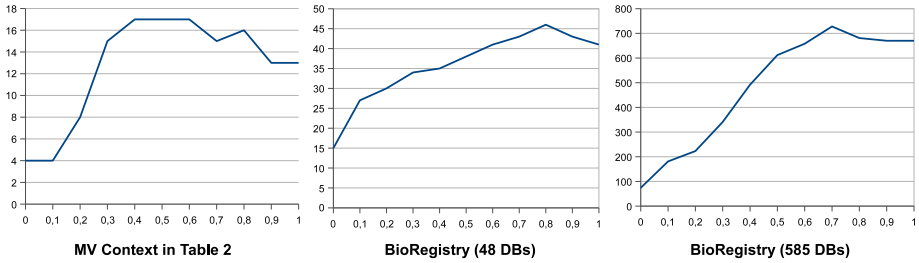

Fig. 5 – Evolution du nombre de concepts multivalués en fonction du seuil de similarité.

5 Conclusion et perspectives

Nous avons introduit, dans cet article, l'ACF par similarité qui s'appuie sur les connaissances de domaine disponibles dans des ressources sémantiques pour étendre les résultats de l'ACF à des données complexes codées sous la forme de contextes multivalués.

5. http ://www.loria.fr/∼messai/SimBAS

L'idée de base est de grouper les objets selon la similarité entre les valeurs des attributs qui les décrivent. La variation de la définition de la similarité permet de faire varier la granularité et le nombre de concepts multivalués dans les treillis obtenus. Cet aspect est utilisé par la suite pour introduire l'exploration progressive, dynamique et interactive des données complexes. Les premières expérimentations de l'approche sur les données de BioRegistry nous ont permis de confirmer les résultats de l'étude théorique de l'ACF par similarité.

La formalisation proposée pour l'ACF par similarité lui permet d'être générique et applicable à différents domaines. Cependant, une application de cette approche à large échelle nécessite une étude algorithmique approfondie à la fois pour construction de treillis et pour le calcul de la similarité entre les valeurs des attributs multivalués. L'ACF par similarité pourra par la suite être présentée pour palier aux limites de l'ACF face à des données complexes dans le cadre d'applications telles que la fouille de données, la recherche d'information, etc.

Références

BELOHLAVEK R. & VYCHODIL V. (2005). What is a fuzzy concept lattice ? In *3rd Internationnal Conference on Concept Lattices and Their Applications, CLA 05, Olomouc, Czech Republic*, p. 34–45.

BRITO P. & POLAILLON G. (2005). Structuring Probabilistic Data by Galois Lattices. *Math. & Sci. hum. / Mathematics and Social Sciences*, **169**(1), 77–104.

CARPINETO C. & ROMANO G. (1996). A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning*, **24**(2), 95–122.

DEVIGNES M.-D., FRANIATTE P., MESSAI N., NAPOLI A. & SMAIL-TABBONE M. (2008). BioRegistry : Automatic Extraction of Metadata for Biological Database Retrieval and Discovery. In *RED'08*, Linz Austria.

FERRÉ S. & RIDOUX O. (2000). A logical generalization of formal concept analysis. In *ICCS'00*, volume 1867 of *LNCS* : Springer.

GANESAN P., GARCIA-MOLINA H. & WIDOM J. (2003). Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems (TOIS)*, **21**(1).

GANTER B. & KUZNETSOV S. O. (2001). Pattern structures and their projections. In *ICCS'01*, volume 2120 of *LNCS* : Springer.

GANTER B. & WILLE R. (1999). *Formal Concept Analysis*. Springer, mathematical foundations edition.

HLIAOUTAKIS A., VARELAS G., VOUTSAKIS E., PETRAKIS E. G. M. & MILIOS E. E. (2006). Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems*, **2**(3), 55–73.

MESSAI N., DEVIGNES M.-D., NAPOLI A. & SMAÏL-TABBONE M. (2008a). Extending attribute dependencies for lattice-based querying and navigation.

MESSAI N., DEVIGNES M.-D., NAPOLI A. & SMAÏL-TABBONE M. (2008b). Many-valued concept lattices for conceptual clustering and information retrieval. In *ECAI'08, 21-25 July, Patras, Greece* : IOS Press.