Différentes notions de réponses pour un système d'interrogation de bases de graphes

Michel Leclère, Nicolas Moreau

▶ **To cite this version:**

Michel Leclère, Nicolas Moreau. Différentes notions de réponses pour un système d'interrogation de bases de graphes. IC 2008 - 19es Journées Francophones d'Ingénierie des Connaissances, Jun 2008, Nancy, France. pp.37-48. hal-00416686

HAL Id: hal-00416686
https://hal.science/hal-00416686

Submitted on 14 Sep 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Différentes notions de réponses pour un système d'interrogation de bases de graphes

Michel Leclère et Nicolas Moreau

LIRMM, Univ. Montpellier 2, CNRS
161, rue Ada
34392 Montpellier, France
{leclere,moreau}@lirmm.fr

Résumé : De nombreuses bases de connaissance sémantiques utilisent des formalismes à base de graphes (graphes conceptuels, RDF, Topic Maps). L'interrogation de telles bases se fonde sur des langages qui pour la plupart expriment par un graphe requête les connaissances recherchées. Nous nous intéressons dans cet article aux réponses exprimées elles-aussi par des graphes et étudions le problème de la redondance entre réponses. Nous différencions les notions de réponse par sous-graphes de la base, de celles de réponse par création de graphes réponses et caractérisons un résultat de requête complet, minimal et sans redondance. Nous définissons également une notion de réponses contextualisées permettant de différencier les réponses grâce à leur voisinage dans la base.

1 Introduction

De nombreuses applications dites "sémantiques" s'appuient sur l'élaboration et l'exploitation de bases de connaissance que ce soit dans le cadre de la gestion documentaire, de la gestion d'un fond audiovisuel, de la construction de mémoires d'entreprise, de l'élaboration d'un portail sémantique, de la gestion de ressources pédagogiques, du web sémantique... Dans de nombreux cas, les formalismes utilisés s'appuient sur des formalismes de graphes pour représenter ces assertions (cf. par exemple les graphes conceptuels, RDF, Topic Maps). L'interrogation de telles bases de connaissance est alors réalisée au moyen de langages qui s'appuient eux-aussi sur le formalisme des graphes (par exemple SPARQL (Prud'hommeaux & Seaborne, 2007)). Une requête est le plus souvent constituée de deux parties : un corps qui permet de spécifier les critères de sélection des pré-réponses (i.e. les parties de la base qui tiennent une réponse à la requête) et une tête qui spécifie comment les réponses sont construites à partir de ces pré-réponses (cf. par exemple les directives SELECT, ASK, CONSTRUCT et DESCRIBE de SPARQL qui spécifient que le format d'une réponse sera un tuple, un booléen ou un graphe). La sélection des pré-réponses est basée sur le calcul des homomorphismes du graphe requête dans la base. Les résultats d'adéquation et complétude de la projection pour les graphes conceptuels ou l'"interpolation lemma" de RDF fondent logiquement

le mécanisme d'interrogation en prouvant qu'on obtient bien toutes les réponses à une requête.

Dans cet article, nous nous focalisons sur les réponses données sous forme de graphes. Il nous semble en effet naturel de conserver un format homogène entre données, requêtes et réponses. Cela permet en particulier de réutiliser les réponses à une requête comme source de données d'une autre requête (cas des requêtes imbriquées). Notons cependant que le problème principal que nous discutons dans cet article, celui de la redondance[1] entre réponses, se pose de la même manière pour des réponses données sous la forme de tuples. De plus, le fait d'utiliser des graphes comme réponses, nous permet d'aborder différemment le problème de la redondance entre réponses puisqu'il nous fournit un moyen de distinguer deux variables par leur voisinage. Enfin, nous avons l'objectif de réaliser un système d'interrogation pour le formalisme des graphes conceptuels et lors de nos recherches sur ce sujet nous n'avons pas trouvé beaucoup de travaux. Les travaux les plus avancés sur cette forme d'interrogation sont ceux de C. Gutierrez et al. (Gutierrez *et al.* (2004)) dans le cadre du web sémantique mais ils ne rentrent pas dans les détails de l'ensemble de "pré-réponses" qu'ils proposent et en particulier n'abordent pas le problème du traitement des redondances entre réponses. SPARQL, quant à lui, propose un modificateur DISTINCT qui élimine les solutions dupliquées mais ne tient pas compte de la redondance.

À titre d'exemple, considérons par exemple la requête et la base de la figure 1 modélisées en graphes conceptuels (cf. section 2). La requête peut être interprétée par : "*quels sont les animaux possédés par une personne ?*" Les parties grisées représentent les 5 pré-réponses.

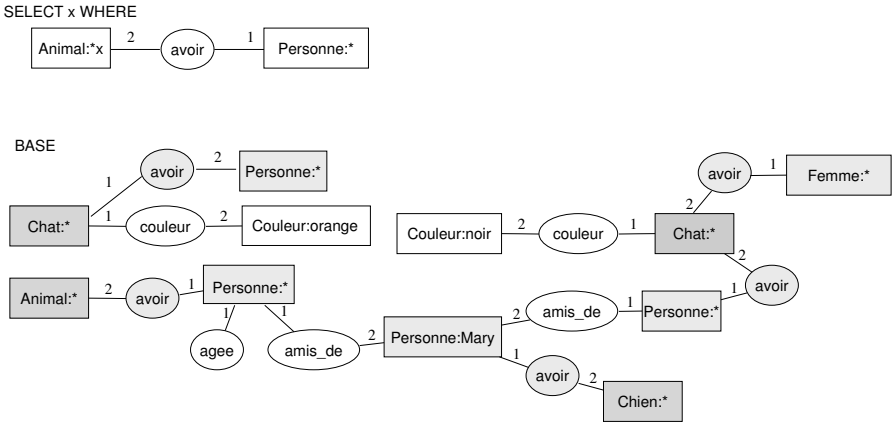


FIG. 1 – Une requête de type SELECT et une base en graphes conceptuels.

Le problème posé est alors de savoir quelles réponses il est pertinent de retourner : si on se contente d'extraire le sommet lié à la variable x dans chaque pré-réponse alors le

[1]Par rapport aux bases de données relationnelles, ces formalismes à base de graphes permettent l'utilisation de variables dans les données et donc dans les réponses, ce qui pose en plus du problème des réponses dupliquées celui de la redondance entre réponses.

résultat de la requête indiquerait : *"il y a un chat", "il y a un chat", "il y a un chat", "il y a un animal", "il y a un chien"*. Observant que deux pré-réponses partagent le même sommet étiqueté `Chat` lié à la variable x, on pourrait éliminer l'une des réponses dupliquées *"il y a un chat"*. Cependant, comme l'hypothèse du monde ouvert est faite sur ces bases de connaissance, il est faux de conclure qu'il y a 4 animaux. Dès lors, il semble préférable de se contenter de donner les réponses "informatives" en préférant le résultat : *"il y a un chat", "il y a un chien"*. Enfin, on pourrait au vu de cette base, considérer qu'un résultat pertinent à cette requête serait : *"il y a un chat de couleur orange", "il y a un chat de couleur noire", "il y a un animal possédé par une personne âgée", "il y a un chien"*.

Nous étudions dans la suite, comment définir ces différents résultats à une requête en nous appuyant sur le formalisme des graphes conceptuels simples (Chein & Mugnier (1992)). Cependant nos résultats peuvent s'appliquer directement à d'autres formalismes de graphes étiquetés en particulier le formalisme RDF/S (Hayes (2004)) des bases de connaissance du web sémantique (cf. Baget (2005) qui montre l'équivalence des deux formalismes). Dans un souci de clarté de la présentation, nous nous limitons à des requêtes "sans tête" en supposant que le format des réponses est simplement celui des pré-reponses.

Nous rappelons dans la section suivante, les principales notions du formalisme des graphes simples sur lequel nous basons notre étude et définissons le modèle d'interrogation considéré. La section 3 introduit différentes notions de réponses en étudiant le problème de la redondance des réponses. Nous différencions les notions de réponse par sous-graphes de la base, de celles de réponse par création de graphes réponses. Sur cette deuxième notion, nous définissons des critères de complétude, non-redondance et minimalité des ensembles de réponses et proposons différentes définitions du résultat d'une requête liées à ces caractéristiques. Dans la section 4, nous introduisons une notion de réponses contextualisées pour différencier les réponses grâce à leur voisinage dans la base afin de limiter la redondance entre réponses.

2 Le modèle d'interrogation étudié

L'étude s'appuie sur le formalisme des graphes conceptuels simples (SG) (cf. Sowa (1984); Chein & Mugnier (1992)). Les graphes simples sont définis sur un support (une ontologie très simple) qui est un quadruplet $S = (T_C,\ T_R,\ I,\ \sigma)$ où T_C est l'ensemble des types de concepts, T_R l'ensemble des types de relations d'arité quelconque (l'arité étant le nombre d'arguments de la relation). T_C et T_R sont partiellement ordonnés. L'ordre partiel représente la relation de spécialisation ($t' \leq t$ est interprété comme "t' est une spécialisation de t"). I est l'ensemble des marqueurs individuels. La relation σ associe à chaque relation une signature définissant son arité et le type maximal de chacun de ses arguments.

Les SG sont des (multi-)graphes bipartis étiquetés notés $G = (C, R, E, l)$ où C et R sont respectivement les ensembles des sommets concepts et relations, E l'ensemble des arêtes, et l la relation d'étiquetage des sommets et des arêtes. Les sommets concepts sont étiquetés par un couple $t : m$ où t est un type de concept et m un marqueur. Lorsqu'un sommet représente une entité non identifiée, son marqueur est générique,

noté $*$, et le sommet est dit *générique*. Autrement, son marqueur est un élément de I, et le sommet est appelé *individuel*. Les sommets relations sont étiquetés par un type de relation r et, si l'on note n l'arité de r, le sommet relation est incident à n arêtes totalement ordonnées (étiquetées de 1 à n). La figure 2 présente des exemples de SG définis sur un support donné.

On appelle *sous-SG* d'un SG $G = (C_G, R_G, E_G, l_G)$, un SG $H = (C_H, R_H, E_H, l_H)$, noté $H \subseteq G$, tel que : $C_H \subseteq C_G$ et $R_H \subseteq R_G$; E_H est la restriction de E_G aux couples de $C_H \times R_H$ et l_H est une restriction de l_G aux éléments de H. Un sous-SG strict d'un SG G est un sous-SG ayant un nombre de sommets strictement plus petit.

Un graphe G est dit *cohérent* par rapport à un support $S = (T_C, T_R, I, \sigma)$ si :
- les étiquettes des sommets concepts (resp. des sommets relations) appartiennent à $(T_C \times (I \cup \{*\}))$ (resp. T_R) ;
- les sommets relations respectent les signatures définies par σ ;
- les types des sommets concepts ayant même marqueur sont comparables deux à deux[2].

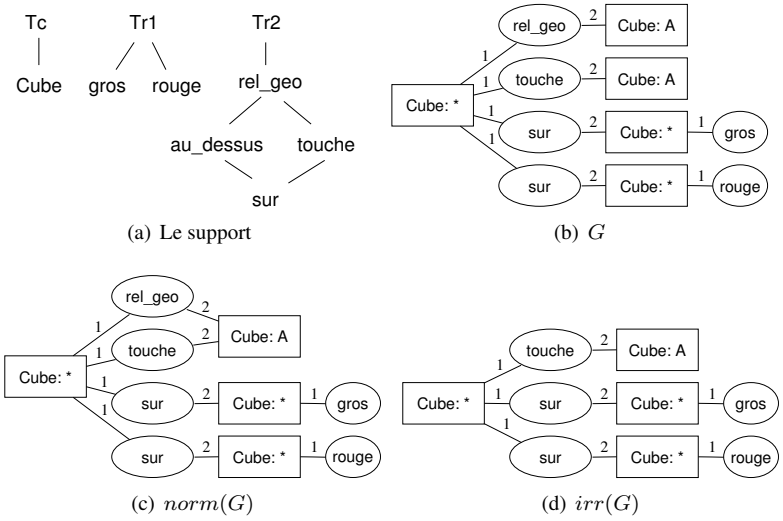
Une relation de spécialisation/généralisation correspondant à une notion de déduction est définie sur les SG et peut être caractérisée par un homomorphisme de graphes appelé *projection*. Lorsqu'il existe une projection π de G dans H, on dit que que H est plus spécialisé que G et l'on note $H \leq G$. Une telle projection est une application de C_G dans C_H et de R_G dans R_H qui préserve les arêtes (si une arête est numérotée i entre r et c dans G, alors il y a une arête numérotée i entre $\pi(r)$ et $\pi(c)$) et peut spécialiser les étiquettes (en respectant l'ordre sur les types et en autorisant la substitution d'un marqueur générique par un marqueur individuel). La figure 1 montre toutes les projections du graphe requête dans le graphe base.

Une sémantique formelle est associée aux SG via une transformation Φ du formalisme SG dans la logique des prédicats. Le résultat fondamental d'*adéquation* et *complétude* établit l'équivalence entre la projection et la déduction sur les formules associées aux SG par Φ : étant donnés deux SG G et H sur un support S, il y a une projection de G dans H si et seulement si $\Phi(S), \Phi(H) \models \Phi(G)$. La complétude est dépendante d'une condition de normalisation sur H : un graphe est sous *forme normale* s'il ne possède pas des sommets concepts différents partageant le même marqueur individuel. Tout graphe cohérent peut être facilement normalisé en fusionnant les sommets concepts ayant même marqueur individuel (et en conservant le type le plus spécifique). On note $norm(G)$ la forme normale d'un graphe G cohérent.

L'équivalence sémantique entre deux graphes G et H, notée $G \equiv H$, peut alors être caractérisée par l'existence d'une projection du premier graphe dans la forme normale de l'autre et du second dans la forme normale du premier. On a ainsi : $\Phi(S) \models \Phi(G) \leftrightarrow \Phi(H)$*ssi* $norm(G) \leq H$ *et* $norm(H) \leq G$. Cette relation d'équivalence définit des classes regroupant des SG équivalents. Au sein de chaque classe d'équivalence, certains graphes contiennent des répétitions inutiles de connaissances (des redondances). Notons qu'un SG qui n'est pas sous forme normale contient des redondances (au moins deux de ses sommets pourraient être fusionnés sans perte d'informations). Un SG est

[2]Cette dernière condition varie selon que l'on considère ou non une relation de conformité dans le support, que l'on impose une structure de treillis à l'ensemble ordonné des types de concept, que l'on introduise des types interdits (i.e. un axiome de disjonction entre types), etc.

dit *redondant* s'il n'est pas sous forme normale ou s'il est équivalent à un de ses sous-graphes stricts. Si ce n'est pas le cas, il est dit *irredondant*[3]. La forme irredondante d'un SG G est notée $irr(G)$. Chaque classe d'équivalence contient un et un seul SG irredondant, qui est le graphe (unique à un isomorphisme près) avec le plus petit nombre de sommets (Mugnier & Chein (1993)). La figure 2 présente G, sa forme normale, et sa forme irredondante.



(a) Le support (b) G

(c) $norm(G)$ (d) $irr(G)$

FIG. 2 – Trois SG équivalents.

Dans la suite, nous considérons un modèle d'interrogation composé d'un *support*, d'un SG *base de connaissance* et d'un SG *requête* (ces SG ne sont pas nécessairement connexes). La base est supposée cohérente par rapport à ce support. Nous ne faisons pas l'hypothèse de l'irredondance de la base de connaissance car d'une part cela ne résoud pas le problème de redondance entre réponses (cf. l'exemple de la figure 1), et d'autre part le calcul de son représentant irredondant serait très coûteux (il est en effet linéaire en la complexité de l'existence d'une projection, qui est lui même un problème NP-complet (Mugnier (1995))). Nous imposons cependant la mise sous forme normale de la base pour assurer la complétude des réponses ; cela est facilement réalisable en un temps linéaire en nombre de sommets de la base, et peut facilement être maintenu incrémentalement. Nous ne faisons aucune hypothèse sur la requête. Il est cependant naturel de vérifier la cohérence du graphe relativement au support de la base pour éliminer les requêtes sans rapport avec la base et on pourrait envisager de calculer la forme irredondante de la requête qui contient généralement peu de sommets.

[3]Notre définition de l'irredondance est plus stricte que celle donnée dans Chein & Mugnier (1992) qui ne prend pas en compte la contrainte de forme normale.

3 Les différentes notions de réponses

Dès lors que l'on considère des formalismes fondés logiquement, on construit les "réponses" à partir des "plus petits sous-graphes" de la base qui ont pour conséquence logique le graphe requête. Pour le formalisme SG, l'existence d'une réponse est donc directement liée à l'existence d'une projection et ces "plus petits sous-graphes" sont tout simplement l'image de la requête par une projection, que nous appellerons *image de preuve* (les projections étant vues comme des preuves de l'existence d'une réponse). Les images de preuves d'une requête Q sur une base B sont donc définies à partir de l'ensemble $\Pi(Q, B)$ des projections de Q dans B. La figure 3 présente l'exemple de la base et de la requête que nous utiliserons dans la cette section. Ces deux graphes sont cohérents par rapport au support de la figure 2(a), et la base correspond à $irr(G)$. Nous avons par ailleurs identifié les sommets des graphes pour pouvoir distinguer les différents sous-graphes. Sur cet exemple, il existe 6 projections : une seule, la projection π_1, est illustrée sur la figure. Nous nous limitons à donner ici les images des sommets relations. Ainsi : $\Pi(Q, B) = \{\pi_1 = \{(r_b, r_1), (r_c, r_2)\}, \pi_2 = \{(r_b, r_1), (r_c, r_3)\}, \pi_3 = \{(r_b, r_2), (r_c, r_2)\}, \pi_4 = \{(r_b, r_3), (r_c, r_3)\}, \pi_5 = \{(r_b, r_2),$
$(r_c, r_3)\}, \pi_6 = \{(r_b, r_3), (r_c, r_2)\}\}$.

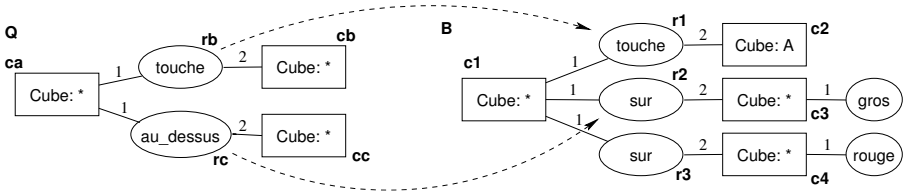


FIG. 3 – Une projection de la requête Q sur la base B.

Dans la suite, nous définissons différentes formes à donner aux graphes réponses en étudiant les redondances possibles entre ces différentes formes.

3.1 Réponse par sous-graphes de la base

La première notion de réponse consiste à retourner les sous-graphes de la base qui sont des images de preuves (il s'agit ici d'identifier des sous-graphes de la base et non de construire des copies de ces sous-graphes). Il peut arriver que deux projections différentes définissent un même graphe image. Nous considérons qu'il s'agit ici de la même réponse (issue de preuves différentes), car distinguer les réponses nécessiterait d'exprimer ces réponses dans un format différent des graphes (par exemple des tuples) permettant d'indiquer les images de chacun des sommets du graphe requête dans la réponse.

Définition 1 (Réponse par sous-graphes images)
On note $R_{IP}(Q, B)$ l'ensemble des images de preuves d'une requête Q dans une base B. $R_{IP}(Q, B) = \{\pi(Q) \mid \pi \in \Pi(Q, B)\}$.

La figure 4 montre les 5 sous-graphes de B répondant à la requête (les graphes images issus des projections π_5 et π_6 correspondant au même sous-graphe de la base, il n'en résulte qu'une seule image de preuve, R_5).

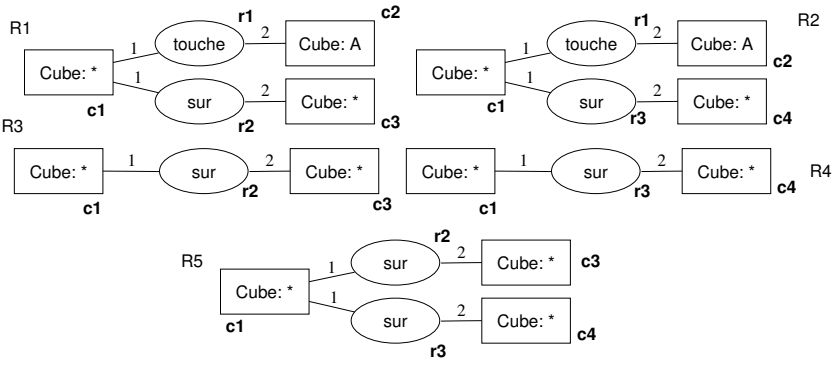


FIG. 4 – Les cinq images de preuves composant $R_{IP}(Q, B)$.

Cette notion de réponses peut être utilisée pour sélectionner des points de départ d'exploration dans la base de connaissance (i.e. un focus sur un sous-graphe de B) que ce soit dans un objectif de "navigation" dans la base de connaissance ou pour servir de base à des requêtes de mise à jour de la base. Dès lors que chacun de ces sous-graphes peut être différencier (par les identifiants des sommets), il n'y a pas de redondance entre ces différentes réponses. Notons que le calcul de cette forme à partir de l'ensemble des images de preuves est linéaire en le nombre de réponses.

3.2 Réponse par graphes indépendants de la base

Dans de nombreux cas, le langage de requête doit permettre "d'extraire" de la connaissance de la base et non "de pointer" des connaissances de la base. Il faut alors que les réponses soient des "copies" de sous-graphes de la base. Les réponses sont alors obtenues par construction de graphes isomorphes aux images de preuves. L'ensemble des réponses n'est plus un ensemble de sous-graphes de B mais un ensemble de graphes isomorphes aux sous-graphes images de preuves de B. Ainsi deux graphes réponses isomorphes doivent être considérés comme égaux. Dans la suite, tous les ensembles de graphes considérés utiliseront cette notion d'égalité entre éléments, ainsi il n'existera aucun graphe isomorphe à un autre graphe dans ces ensembles.

Définition 2 (Iso-Réponse)
Une iso-réponse de Q dans B est un graphe isomorphe à une image de preuve de Q dans B. On note $R_{ISO}(Q, B)$, l'ensemble des iso-réponses $R_{ISO}(Q, B) = \{G \mid$ il existe $G_i \in R_{IP}(Q, B)$ avec G_i isomorphe à $G\}$

La figure 5 montre les graphes composant R_{ISO} dans notre exemple. Dès lors que nous imposons à nos bases d'être sous forme normale, les graphes réponses sont eux-aussi forcément sous forme normale. Il n'en est pas de même pour la propriété d'irre-

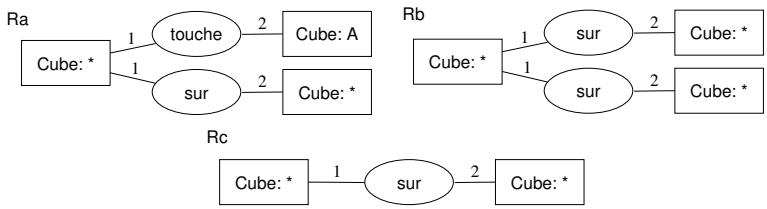


FIG. 5 – Les graphes composant $R_{ISO}(Q, B)$.

dondance : la réponse R_b est redondante bien que la requête et la base soient irredondantes.

Avec cette notion de réponse, on perd le lien avec la base, seule l'attestation de l'existence d'une connaissance particulière dans la base est conservée. On peut dès lors s'interroger sur la pertinence de ne considérer qu'une équivalence syntaxique (l'isomorphisme) comme critère d'équivalence de deux réponses, et non l'équivalence sémantique : quel intérêt y-a t'il à donner la réponse R_b quand on a la réponse R_c ? En effet, $R_c = irr(R_b)$.

Il nous paraît même souhaitable de ne conserver que des réponses amenant une connaissance nouvelle par rapport aux autres réponses. Si l'on considère l'exemple introductif (cf. figure 1), quel intérêt y-a t'il à donner la réponse `[Animal:*]` quand on a la réponse `[Chat:*]` ? Pour cela nous introduisons un critère de comparabilité des réponses : deux réponses sont comparables si les connaissances assertées par l'une sont déductibles de celles assertées par l'autre. Dans ce cas, nous considérons que la première est redondante par rapport à la seconde. L'idée est de ne renvoyer que des réponses incomparables.

Définition 3 (Critère d'incomparabilité)
Un ensemble de réponses $R(Q, B) \subseteq R_{ISO}(Q, B)$ est dit "sans redondance" lorsque ses réponses sont incomparables i.e. $\forall G_i, G_j \in R(Q, B)$ avec $G_j \neq G_i$, on a $G_i \not\leq G_j$.

Par ailleurs, une propriété souhaitable est que l'ensemble des réponses retournées exprime la connaissance contenue dans toutes les iso-réponses (on cherche à éliminer des redondances mais pas à perdre de la connaissance). En considérant que l'interprétation logique d'un ensemble de réponse est la conjonction des formules associées à chaque réponse (i.e. si $R = \{r_1, ...r_n\}$, $\Phi(R) = \Phi(r_1) \wedge ... \wedge \Phi(r_n)$), on introduit le critère de complétude suivant.

Définition 4 (Critère de complétude)
Un ensemble de réponses $R(Q, B) \subseteq R_{ISO}(Q, B)$ est dit "complet" ssi $\Phi(S), \Phi(R(Q, B)) \models \Phi(R_{ISO}(Q, B))$.

Enfin, il semble naturel de choisir l'ensemble de réponses le plus petit. Cela amène à définir un critère de minimalité qui s'appuie sur une notion de *taille de réponse* et d'*équivalence de deux ensembles de réponses*.

Définition 5 (Critère de minimalité)
La taille d'un ensemble réponse est la somme du nombre de sommets des graphes composant l'ensemble. Deux ensembles de réponses sont équivalents ssi $\Phi(S) \models \Phi(R(Q,B)) \leftrightarrow \Phi(R'(Q,B))$. *Un ensemble de réponses* $R(Q,B) \subseteq R_{ISO}(Q,B)$ *est minimal ssi il n'existe pas un ensemble équivalent de réponses de taille strictement inférieure.*

Le critère de minimalité amène à choisir la forme irredondante des réponses comme représentant unique de chaque classe d'équivalence (puisqu'elle est minimale). La propriété 1 atteste que toute forme irredondante d'une réponse est également une réponse.

Propriété 1
Quelque soit Q *et* B, *pour tout* $R \in R_{IP}(Q,B)$ *on a* $irr(R) \in R_{IP}(Q,B)$

Dès lors qu'on couple la minimalité au critère de complétude cela permet de définir une notion de réponse appropriée pour une interrogation visant à retrouver *toutes les connaissances* d'une base répondant à une requête.

Définition 6 (Réponses les plus spécifiques)
On note $R_{MIN}(Q,B)$ *l'ensemble des réponses irredondantes les plus spécifiques :* $R_{MIN}(Q,B) = \{irr(G) \in R_{ISO}(Q,B) \mid \nexists G' \in R_{ISO}(Q,B)$ *avec* $G' < G\}$.

Sur l'exemple de la figure 5, $R_{MIN}(Q,B) = \{R_a\}$ (sur l'exemple introductif, on aurait $\{$`[Chien:*],[Chat:*]`$\}$). $R_{MIN}(Q,B)$ cumule les propriétés d'incomparabilité, de minimalité et de complétude. Il est en outre le seul ensemble à avoir toutes ces propriétés (cf. théorème 1). Le calcul de cette forme de réponse est quadratique en le nombre d'appels au problème NP-complet de l'existence d'une projection ; cependant les instances sur lesquelles ce problème est exécuté sont relativement petites.

Théorème 1
$R_{MIN}(Q,B)$ *est l'unique sous-ensemble de réponses incomparables minimal complet de* $R_{ISO}(Q,B)$.

Preuve : • Incomparabilité : $R_{MIN}(Q,B)$ est par définition composé des éléments irredondants minimaux de $R_{ISO}(Q,B)$ et comme par ailleurs deux éléments irredondants et non isomorphes ne sont pas équivalents, toutes les réponses de $R_{MIN}(Q,B)$ sont bien incomparables. • Complétude : La forme irredondante d'une réponse étant elle même une réponse, l'ensemble des réponses irredondantes est complet. Par définition de $R_{MIN}(Q,B)$, on ne garde de cet ensemble que les G_i qui sont plus spécifiques qu'un autre graphe de $R_{ISO}(Q,B)$. $R_{MIN}(Q,B)$ est donc complet. • Minimalité et Unicité : $R_{MIN}(Q,B)$ étant composé uniquement de graphes irredondants, il est minimal et unique (à un isomorphisme près).

4 Les réponses contextualisées

Le critère de complétude des réponses, défini précédemment, assure qu'aucune information n'est perdue lors de la suppression d'une réponse, qui est alors jugée redondante (au sens de la comparabilité) vis-à-vis d'une autre réponse. Or cette comparabilité

n'est exprimée que sur la seule connaissance des images de preuves. La diversité des images de preuves (l'ensemble potentiel d'images de preuves incomparables) est limitée par la taille de la requête, mais également par leur propriété intrinsèque d'être toutes des spécialisations d'un même graphe requête. Pour augmenter cette diversité, une idée consiste à contextualiser les réponses en leur ajoutant suffisament de connaissance (prise dans la base) pour les rendre incomparables. Nous étudions dans cette section les avantages et les limites de ce processus de contextualisation des réponses. Notons que ce mécanisme de contextualisation est proposé dans SPARQL (cf. les requêtes DESCRIBE) dans un objectif purement descriptif et non pour traiter le problème de redondance entre réponses.

Contextualiser consiste à ajouter de la connaissance aux différentes images de preuves. Une image de preuve contextualisée est un graphe isomorphe à un sous-graphe de la base contenant une ou plusieurs images de preuves. Nous imposons qu'une image de preuve n'apparaisse que dans une seule image de preuve contextualisée, pour ne pas dupliquer les réponses. Nous nous limitons (dans cette première étude) à considérer des contextualisations qui ne regroupent pas les images de preuves. La seule exception concerne des images de preuves incluses (i.e. l'une est sous-graphe de l'autre). C'est le cas par exemple des sous-graphes R_4 et R_2 de la figure 4. Dans ce cas, la contextualisation du sous-graphe est celle du sur-graphe (c'est à dire qu'il n'existera qu'une seule image de preuve contextualisée, regroupant les deux images de preuves).

Définition 7 (Contextualisation)
Une contextualisation C est une application qui à un ensemble d'images de preuves $R_{IP}(Q, B)$ associe un ensemble d'images de preuves contextualisées $R^C_{IP}(Q, B) = \{C(x) | x \subseteq C(x) \subseteq B$ et $x \in R_{IP}(Q, B)\}$ telle que

- *chaque image de preuve x n'est sous-graphe que d'une seule image de preuve contextualisée, la sienne $C(x)$;*
- *et soient deux images de preuves x et y, $C(x) = C(y)$ ssi $x \subseteq y$ ou $y \subseteq x$.*

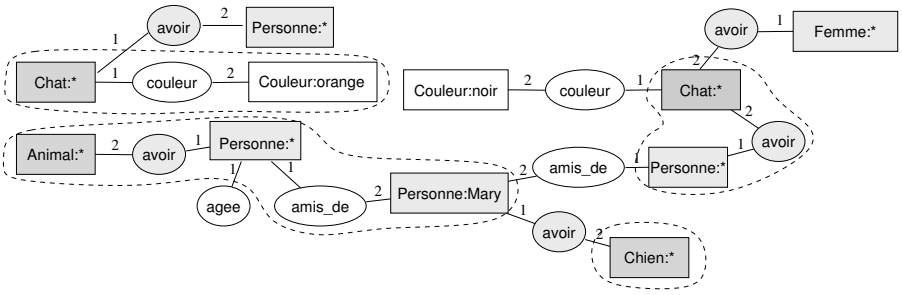


FIG. 6 – Un exemple de contextualisation.

La figure 6, qui reprend l'exemple de la figure 1, propose une contextualisation (en pointillés) des images de preuves permettant de différencier toutes les réponses. Ainsi on aurait bien quatre réponses au lieu des deux fournies par la notion de réponses les plus spécifiques (R_{MIN}).

Un critère à imposer à une contextualisation est son honnêteté : deux réponses rendues incomparables doivent l'être pour de vraies raisons. Une contextualisation est honnête sur une base et une requête données si pour toute image de preuve, lorsque la connaissance exprimée par son contexte (la partie ajoutée) est présente dans le voisinage d'une autre image de preuve, alors le contexte de cette dernière contient aussi cette connaissance. Dans l'exemple de la figure 6, la contextualisation n'est pas honnête, car l'animal est rendu incomparable avec un des chats car "il appartient à une personne amie de Mary", or cette information est également vraie pour le chat en question. Le même problème se trouve également entre les deux chats, l'un étant différencié par une connaissance ("il appartient à une personne") commune aux deux chats.

Définition 8 (Honnêteté)
Une contextualisation C est honnête sur (Q, B) si pour tout $x \in R_{IP}(Q, B)$, s'il existe une projection π (différente de l'identité) de $C(x)$ dans B (on note $\pi(x) = y \in R_{IP}(Q, B)$) alors $C(y) \leq C(x)$.

Pour un ensemble d'images de preuves donné, il peut exister de nombreuses contextualisations honnêtes. Nous proposons ici une contextualisation qui s'appuie sur la distance k des sommets voisins de ceux de l'image de preuve.

Définition 9 (Voisinage à distance k)
Étant donné $R \in R_{IP}(Q, B)$, le voisinage à distance k de R est le graphe contenant R augmenté de tous les sommets concepts pouvant être atteints par un nombre de relations $\leq k$ à partir d'un concept de R, ainsi que toutes ces relations.

Nous définissons notre contextualisation incrémentalement en choisissant le plus petit k tel que l'ensemble des voisinages à distance k, V_k, des images de preuves soit une contextualisation, et qu'il n'existe pas un $k' \geq k$ tel que l'ensemble des voisinages à distance k', $V_{k'}$, soit une contextualisation telle que $|V_{k'}| > |V_k|$ ($V_{k'}$ permettrait d'avoir plus de réponses que V_k).

La méthode de construction de notre contextualisation à distance k assure que cette contextualisation sera toujours honnête quelque soit la base, la requête et la distance considérées. Cette propriété vient de la "conservation" des distances par la projection : si la connaissance d'un contexte de distance k est exprimée dans le voisinage d'une autre image de preuve, alors cette connaissance se trouve forcément dans le voisinage de distance $\leq k$ de cette image de preuve. La figure 7 montre la contextualisation à distance 2. La contextualisation de distance 1 ne différenciait pas toutes les images de preuves (par exemple l'animal et le chien). À distance 2, toutes les réponses sont incomparables, ce sera donc la distance choisie.

5 Conclusion

Dans le cadre de l'interrogation de bases de connaissance, nous avons défini deux principales notions de réponses à une requête : la première étant composée de sous-graphes de la base permet de démarrer des navigations dans la base de connaissance ou de servir de base à des requêtes de mises à jour ; la seconde étant constituée de

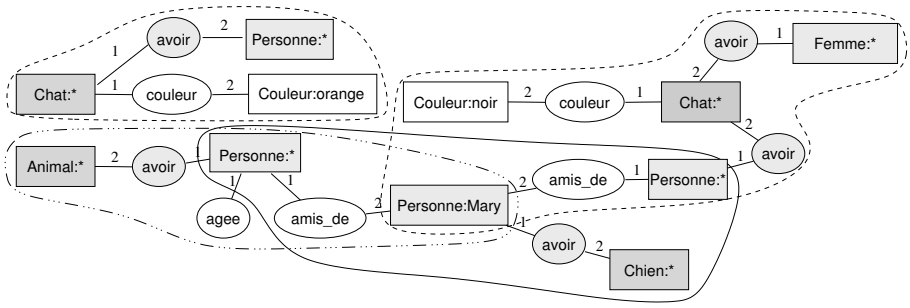


FIG. 7 – Contextualisation à distance 2.

graphes indépendants de la base. Nous avons défini plusieurs critères désirables pour cette dernière notion : l'incomparabilité des réponses, la complétude de l'ensemble des réponses et la minimalité (en taille) de l'ensemble de réponses. La notion de réponse qui semble la plus intéressante en regard de ces critères est celle des réponses irredondantes les plus spécifiques.

Nous proposons une notion de réponses contextualisées pour permettre de différencier les réponses et ainsi éliminer des redondances sans perte de réponses. Nous avons défini un critère d'honnêteté pour une telle contextualisation, et proposé une contextualisation basée sur le voisinage d'une image de preuve à distance k.

Remerciements

Ce travail est financé par le programme ANR-RNTL dans le cadre du projet Eiffel.

Références

BAGET J.-F. (2005). Rdf entailment as a graph homomorphism. *The Semantic Web : ISWC 2005*, p. 82–96.

CHEIN M. & MUGNIER M.-L. (1992). Conceptual Graphs : Fundamental Notions. *Revue d'Intelligence Artificielle*, **6**(4), 365–406.

GUTIERREZ C., HURTADO C. & MENDELZON A. O. (2004). Foundations of semantic web databases. In *PODS '04 : Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, p. 95–106, New York, NY, USA : ACM Press.

HAYES P. (2004). *RDF Semantics*. Rapport interne, W3C.

MUGNIER M. (1995). On generalization/specialization for conceptual graphs. *Journal of Experimental & Theoretical Artificial Intelligence*, **7**(3), 325–344.

MUGNIER M.-L. & CHEIN M. (1993). Polynomial algorithms for projection and matching. In *Proceedings of the 7th Annual Workshop on Conceptual Structures : Theory and Implementation*, p. 239–251, London, UK : Springer-Verlag.

PRUD'HOMMEAUX E. & SEABORNE A. (2007). *SPARQL Query Language for RDF*. Rapport interne, W3C.

SOWA J. F. (1984). *Conceptual Structures : Information Processing in Mind and Machine*. Addison-Wesley.