Interrogation par analogie dans les bases de données

Sara El Hassad

Interrogation par analogie dans les bases de données

Sara El Hassad
Université de Rennes 1 – IRISA/SHAMAN
sara.el-hassad@irisa.fr
Ce travail s'inscrit dans le projet PAWS partiellement financé par
Lannion-Trégor Communauté et la Région Bretagne

1. CONTEXTE ET MOTIVATIONS

L'*analogie*, notion bien connue en Sciences Cognitives et en Intelligence Artificielle, met en relation deux objets A et B sous la forme d'une assertion : « *A est comme B* » [7]. Elle se distingue de la similarité en mettant en avant les caractéristiques communes de deux objets, a priori différents, sur la base de leurs définitions. Ainsi, selon [7], un soleil est comme un noyau d'atomes car des objets gravitent autour d'eux (resp., des planètes et des électrons) : ce sont des centres de systèmes gravitationnels. La *proportion analogique* caractérise une analogie des relations entre une paire d'objets (A, B) d'une part, et entre une paire d'objets (C, D) d'autre part, sous la forme d'une assertion : « *A est à B ce que C est à D* » [10]. De l'analogie précédente, on tire la proportion analogique suivante : une planète est à un soleil ce qu'un électron est à un noyau d'atome.

L'objectif général de la thèse est d'étudier l'interrogation par analogie des bases de données. Nos contributions en cours et à venir sont la formalisation de ce nouveau type d'interrogation, l'énumération des problèmes de décision et de calcul associés et la résolution des challenges algorithmiques qui en découlent. Ces travaux doivent par exemple permetre à un journaliste de données, face à des données d'observations des écosystèmes naturels nationaux, de rechercher des analogies entre des espèces différentes. Celui-ci pourrait être amené à formuler des interrogations fondées sur l'analogie telles que : « Quelles espèces sont comme les algues vertes observées sur les côtes bretonnes ? » ; ou basées sur la proportion analogique liant les espèces à leur lieu de vie : « Quelles espèces sont aux rivières d'Aquitaine ce que les algues vertes sont aux côtes bretonnes ? »

La suite de cet article est organisée comme suit. Dans la section 2, nous formalisons ces relations en termes de notions classiques de Bases de Données et nous les illustrons. Dans la section 3, nous exposons les différents problèmes de décision et de calcul que nous allons étudier. Enfin, dans la section 4, nous pointons des travaux pouvant servir de bases à l'élaboration d'algorithmes pour résoudre ces problèmes.

BDA 29 Septembre-02 Octobre 2015, Ile de Porquerolles, France.

2. ANALOGIE ET BASES DE DONNÉES

Nous formalisons ici les relations d'analogie et de proportion analogique dans un cadre général de bases de données.

Étant donné un alphabet de relations $\mathcal{R}$, une base de données est l'union d'un ensemble de contraintes $\mathcal{C}$ sur $\mathcal{R}$ et d'un ensemble de faits $\mathcal{F}$ sur $\mathcal{R}$. Ainsi, nous capturons le cadre des bases de données relationnelles lorsque $\mathcal{C}$ modélise des contraintes de dépendance fonctionnelle, d'inclusion, etc, ainsi que celui des bases de données ontologiques (ou bases de connaissances) quand $\mathcal{C}$ modélise des contraintes déductives à l'aide d'un dialect de logiques de description [2], de Datalog± [4] ou encore de Règles Existentielles [3]. Les objets impliqués dans les relations d'analogie et de proportion analogique sont des requêtes conjonctives sur $\mathcal{R}$.

Nous considérons deux objets comme étant en analogie, s'ils appartiennent à un même type plus général que le leur. Cette relation s'exprime naturellement dans notre cadre avec la notion d'*inclusion de requêtes* [1] : étant donné $\mathcal{C}$ sur $\mathcal{R}$, une requête Q' est incluse dans une requête Q, noté $Q' \subseteq_{\mathcal{C}} Q$, si et seulement si toutes les réponses de Q' sont des réponses de Q, pour tout $\mathcal{F}$ sur $\mathcal{R}$.

Définition 1 (Analogie). *Soit Q_A et Q_B deux requêtes sur $\mathcal{R}$. Q_A* est comme *Q_B ssi il existe une requête Q sur $\mathcal{R}$ telle que $Q_A \subseteq_{\mathcal{C}} Q$ et $Q_B \subseteq_{\mathcal{C}} Q$, où $\subseteq_{\mathcal{C}}$ est l'inclusion de requêtes modulo $\mathcal{C}$.*

Nous considérons que deux paires d'objets, (A, B) et (C, D), sont en proportion analogique si une relation entre A et B *est comme* une relation entre C et D. Ces relations entre objets s'expriment naturellement par des requêtes et leur analogie se traduit encore à l'aide de l'inclusion de requêtes.

Définition 2 (Proportion Analogique). *Soit Q_A, Q_B, Q_C et Q_D des requêtes sur $\mathcal{R}$. Q_A* est à *Q_B* ce que *Q_C* est à *Q_D ssi :*

- *il existe une requête Q_{AB} sur $\mathcal{R}$ telle que $Q_{AB} \not\subseteq_{\mathcal{C}} \perp$ et $Q_{AB} \subset_{\mathcal{C}} Q_A \times Q_B$,*
- *il existe une requête Q_{CD} sur $\mathcal{R}$ telle que $Q_{CD} \not\subseteq_{\mathcal{C}} \perp$ et $Q_{CD} \subset_{\mathcal{C}} Q_C \times Q_D$,*
- *il existe une requête Q sur $\mathcal{R}$ telle que $Q_{AB} \subseteq_{\mathcal{C}} Q$ et $Q_{CD} \subseteq_{\mathcal{C}} Q$.*

On remarquera que la définition ci-dessus impose aux requêtes exprimant des relations entre objets d'être *non-vides* et *sans* produit cartésien, afin de modéliser un lien sémantique réel entre les objets liés : sans cette contrainte, tout quadruplet d'objets formerait une proportion analogique.

Exemple (**Biodiversité**) Soit l'ensemble de contraintes $\mathcal{C} = \{\forall x(\texttt{AlgueVerte}(x) \Rightarrow \texttt{Espece}(x)),$

$\forall x(\texttt{Ecrevisse}(x) \Rightarrow \texttt{Espece}(x))$,
$\forall x(\texttt{AlgueVerte}(x) \wedge \texttt{Ecrevisse}(x) \Rightarrow \bot)$,
$\forall x(\texttt{BaieBretonne}(x) \Rightarrow \texttt{ZoneObservation}(x))$,
$\forall x(\texttt{RiviereAquitaine}(x) \Rightarrow \texttt{ZoneObservation}(x))$,
$\forall x \forall y(\texttt{prolifere}(x,y) \Rightarrow \texttt{envahit}(x,y))$,
$\forall x \forall y(\texttt{contamine}(x,y) \Rightarrow \texttt{envahit}(x,y))$,
$\forall x \forall y(\texttt{prolifere}(x,y) \Rightarrow \texttt{AlgueVerte}(x))$,
$\forall x \forall y(\texttt{prolifere}(x,y) \Rightarrow \texttt{BaieBretonne}(y))$,
$\forall x \forall y(\texttt{contamine}(x,y) \Rightarrow \texttt{Ecrevisse}(x))$,
$\forall x \forall y(\texttt{contamine}(x,y) \Rightarrow \texttt{RiviereAquitaine}(y))\}$.

Les objets $Q_A(x)$:−$\texttt{AlgueVerte}(x) \wedge \texttt{prolifere}(x,y) \wedge \texttt{BaieBretonne}(y)$ et $Q_B(x)$:−$\texttt{Ecrevisse}(x) \wedge \texttt{contamine}(x,y) \wedge \texttt{RiviereAquitaine}(y)$ sont en analogie car la requête $Q(x)$: − $\texttt{Espece}(x) \wedge \texttt{envahit}(x,y) \wedge \texttt{ZoneObservation}(y)$ les inclut.

Les objets $Q_A(x)$:−$\texttt{AlgueVerte}(x)$, $Q_B(x)$:−$\texttt{BaieBretonne}(x)$, $Q_C(x)$:−$\texttt{Ecrevisse}(x)$ et $Q_D(x)$:−$\texttt{RiviereAquitaine}(x)$ sont en proportion analogique car il existe $Q_{AB}(x,y)$:−$\texttt{AlgueVerte}(x) \wedge \texttt{prolifere}(x,y) \wedge \texttt{BaieBretonne}(y)$ qui lie les algues vertes aux baies Bretonnes, $Q_{CD}(x,y)$:−$\texttt{Ecrevisse}(x) \wedge \texttt{contamine}(x,y) \wedge \texttt{RiviereAquitaine}(y)$ qui lie les écrevisses américaines aux rivières d'Aquitaine et $Q(x,y)$:−$\texttt{Espèce}(x) \wedge \texttt{envahit}(x,y) \wedge \texttt{ZoneObservation}(y)$ qui les inclut. ⋄

3. PROBLÈMES À ÉTUDIER

Sur la base du cadre défini ci-dessus, nous nous intéressons à l'algorithmique de problèmes de décision et de calcul.

Les problèmes de décision consistent, étant donné un ensemble de contraintes $\mathcal{C}$, à dire – Oui ou Non – si deux objets sont en analogie ou si deux paires d'objets sont en proportion analogique. Ceci revient à *décider* s'il existe une requête Q au sens de la Définition 1 pour l'analogie ou de la Définition 2 pour la proportion analogique. Les problèmes de calcul consistent (i) à expliquer et (ii) à découvrir des analogies ou des proportion analogiques. Expliquer une analogie revient, pour une instance positive d'un problème de décision décrit ci-dessus, à *exhiber* une requête Q susmentionnée. Découvrir une analogie ou proportion analogique consiste à trouver les *instances positives* de problèmes de décision décrits ci-dessus, dont les objets sont partiellement connus. Ceci permet de poser des questions telles que : Quel x est comme Q_B ? Quel x est à Q_B ce que Q_C est à Q_D ? Etc.

Les challenges algorithmiques sous-jacents à la résolution de ces problèmes sont doubles. Tout d'abord, il s'agit de *calculer* des généralisants communs de requêtes modulo un ensemble de contraintes, en particulier des plus petits au regard de la relation d'inclusion de requêtes. Les généralisants communs permettent de caractériser l'existence de requêtes Q au sens de la Définition 1 et de la Définition 2 ; les plus petits généralisants communs permettent de les expliquer de façon la plus intelligible possible à un utilisateur, car ils sont plus proches sémantiquement des l'énoncés des problèmes. Pour calculer les requêtes Q_{AB} et Q_{CD} au sens de la Définition 2, nous faisons appel à la spécialisation de requêtes modulo un ensemble de contraintes. Ensuite, pour les problèmes autres que la décision et l'explication d'analogie, il s'agit de chercher des requêtes (objets ou relations entre eux) caractérisant une analogie ou proportion analogique. On notera que l'espace de recherche est particulièrement grand. Par exemple, dans le cas le plus simple, chercher des x comme un Q_A donné, consiste à trouver toutes les requêtes incluses dans toutes celles généralisant Q_A ! Il faudra donc restreindre l'espace effectif de recherche avec des métriques sur la relation d'inclusion éventuellement complétée par des statistiques sur les données interrogées, afin d'éviter les généralisations et spécialisations excessives.

4. TRAVAUX EN COURS

Notre cadre d'interrogation par analogie pour les bases de données ontologiques utilise la logique de description « légère » DL-lite$_\mathcal{R}$ ([5]) comme langage de contraintes. Ce fragment pragmatique de Datalog± et des Règles Existentielles fournit les fondements d'OWL2 QL, un modèle de données sémantiquement riche du W3C. Actuellement, nous nous focalisons sur le calcul, fondamental à la mise en place du cadre d'interrogation proposé, de généralisants et en particulier de plus petits généralisants communs. À notre connaissance, ce problème n'a pas été étudié pour DL-lite$_\mathcal{R}$, contrairement au problème dual de calcul de requêtes incluses dans une requête donnée qui est une composante nécessaire pour répondre à une requête par reformulation (par ex. [5, 8]). Nous nous inspirons donc de résultats algorithmiques issus de différents domaines de l'informatique dans lesquels le problème du calcul d'un plus petit généralisant commun a été traité pour des formalismes logiques proches. En particulier, ce problème admet une solution (trop) générale pour les *clauses de Horn sans symbole de fonction* en Programmation Logique Inductive [9], ainsi qu'une solution (trop) spécifique pour les *graphes conceptuels simples* en Représentation des Connaissances [6]. Nos algorithmes se fonderont sur ces travaux afin de proposer une solution adaptée et optimisée pour DL-lite$_\mathcal{R}$.

5. REFERENCES

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge Univ. Press, 2003.

[3] J. Baget, M. Leclère, M. Mugnier, and E. Salvat. On rules with existential variables : Walking the decidability line. *Artificial Intelligence*, 175, 2011.

[4] A. Calì, G. Gottlob, and T. Lukasiewicz. A general datalog-based framework for tractable query answering over ontologies. In *PODS*, 2009.

[5] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics : The *DL-Lite* family. *JAR*, 39(3) :385–429, 2007.

[6] M. Chein and M.-L. Mugnier. *Graph-based Knowledge Representation : Computational Foundations of Conceptual Graphs*. Springer, 2008.

[7] D. Gentner. Structure-mapping : A theoretical framework for analogy. *Cognitive Science*, 7(2), 1983.

[8] G. Gottlob, G. Orsi, and A. Pieris. Query rewriting and optimization for ontological databases. *ACM TODS*, 39(3), 2014.

[9] S.-H. Nienhuys-Cheng and R. de Wolf. Least generalizations and greatest specializations of sets of clauses. *J. of Artificial Intelligence Research*, 1996.

[10] H. Prade and G. Richard, editors. *Computational Approaches to Analogical Reasoning : Current Trends*, volume 548 of *Studies in Computational Intelligence*. Springer, 2014.