Analyse de concepts formels pour la classification de triplets RDF

Justine Reynaud, Yannick Toussaint, Amedeo Napoli

Analyse de concepts formels pour la classification de triplets RDF

Justine Reynaud
LORIA
(CNRS, INRIA, U. Lorraine)
BP 239
54506 Vandoeuvre-les-Nancy
justine.reynaud@loria.fr

Yannick Toussaint
LORIA
(CNRS - INRIA - U. Lorraine)
BP 239
54506 Vandoeuvre-les-Nancy
yannick.toussaint@loria.fr

Amedeo Napoli
LORIA
(CNRS - INRIA - U. Lorraine)
BP 239
54506 Vandoeuvre-les-Nancy
amedeo.napoli@loria.fr

ABSTRACT

L'émergence du web des données et des *linked open data* donne lieu à de nouvelles problématiques d'abstraction, d'organisation et d'interprétation. Nous nous intéressons ici à la classification de triplets `RDF`. Pour cela, nous nous appuyons sur l'analyse de concepts formels (FCA) et ses extensions, notamment les structures de patrons. La FCA permet notamment de construire une représentation visuelle de la classification et permet à l'utilisateur de naviguer facilement au sein de celle-ci.

1. INTRODUCTION

Le web des données a permis la création de nombreuses bases de connaissances librement accessibles en ligne et liées les unes au autres. DBPedia est l'une d'entre elles. L'objectif de DBPedia est de représenter sous forme structurée les connaissances disponibles sur Wikipedia. Cette structure s'appuie sur des langages standards tels que `RDF`, `RDFS` et `OWL`.

On considère un ensemble de ressources sémantiquement liées entre elles. Différents types de ressources sont distingués : les *prédicats* correspondent à une relation binaire entre deux ressources, les *classes* représentent des ensembles de ressources, et les *instances* représentent les ressources qui ne sont ni des prédicats ni des classes. L'unité de base pour exprimer une connaissance est le triplet (sujet, prédicat, objet). Il peut exprimer des faits (`B_Obama néA Honolulu`), mais aussi structurer les ressources entre elles, notamment grâce aux prédicats `rdfs:subClassOf` et `rdfs:subPropertyOf` (notées par la suite `subC` et `subP` respectivement). Par exemple, `Capitale rdfs:subClassOf Ville` indique que toutes les capitales sont aussi des villes. Un exemple d'ensemble de triplets ainsi que les hiérarchies de classes et de prédicats associés sont présentés figure 1.

Ici, nous souhaitons classifier un ensemble de triplets en tenant compte des hiérarchies de classes et de prédicats impliqués. Nous considérons donc d'une part les ensembles de triplets à classifier, et de l'autre les hiérarchies.

BDA 2016, 15 au 18 Novembre, Poitiers, France.

2. APPROCHES EXISTANTES

Des travaux existent dans le domaine du web sémantique (par exemple, typer automatiquement les ressources [9]) et dans le domaine des logiques de description notamment. Certain d'entre eux s'appuient sur la FCA [10].

La FCA [7] est une théorie mathématique qui permet de classifier des objets en fonction des attributs qu'ils partagent. Étant donné un ensemble G d'objets, un ensemble M d'attributs, et une relation binaire $I \subseteq G \times M$, $\mathbb{K} = (G, M, I)$ est un contexte formel. gIm s'interprète comme *g a pour attribut m*. Les correspondances de Galois (notées $.'$) sont définies comme suit : $A' = \{m \mid \forall a \in A, aIm\}$ avec $A \subseteq G$ et $B' = \{g \mid \forall b \in B, gIb\}$ avec $B \subseteq M$. Elles permettent de définir un concept formel (A, B) comme un sous-ensemble d'objets ($A \subseteq G$) partageant un sous-ensemble maximal d'attributs ($B \subseteq M$) et qui vérifie $A'' = A$ et $B'' = B$. A et B sont appelés *extent* et *intent* respectivement.

De nombreuses extensions ont été proposées afin de prendre en compte des données plus complexes. Dans [4], les auteurs proposent de considérer une *background knowledge*, une connaissance extérieure au contexte, en tenant compte d'une hiérarchie entre les attributs.

Les structures de patrons [6] sont une autre extension permettant de tenir compte de la hiérarchie entre les attributs. Au lieu de considérer un ensemble d'attributs, on considère un ensemble de descriptions partiellement ordonnées $(D, \sqcap)$. Chaque objet est associé à une description. À la place d'une intersection entre les ensembles d'attributs, une opération de similarité entre les descriptions est définie. Un concept a donc pour extent un ensemble d'objets et pour intent une description. Dans [3], une classification des sujets de chaque triplet reposant sur une structure de patrons est proposée. Étant donné un ensemble de triplets $\mathcal{B}$, les auteurs considèrent la description d'un sujet s comme l'ensemble des couples $(p, C(o))$ tels que $(s, p, o) \in \mathcal{B}$, où $C(o)$ représente la classe de o. La similarité entre deux sujets s_1 et s_2 est l'ensemble des couples (p, C) de façon à ce qu'il existe x, y tels que $(s_1, p, x), (s_2, p, y) \in \mathcal{B}$ et $C(x)$ `subC` C, $C(y)$ `subC` C. Cette approche s'appuie sur la hiérarchie des classes considérées, mais ne prends donc pas en compte la hiérarchie de prédicats. D'autres approches basées sur la FCA ont été proposées, en s'appuyant sur une représentation sous forme de graphes, notamment dans [5] et [8].

3. APPROCHE PROPOSÉE

Nous proposons une structure de patrons permettant de classifier les triplets en tenant compte des hiérarchies de

classes et de prédicats.

3.1 Définition

Étant donné un triplet (s, p, o), sa description est $\delta((s, p, o)) = (C(s), p, C(o))$. La similarité entre deux triplets $t_1 = (s_1, p_1, o_1)$ et $t_2 = (s_2, p_2, o_2)$ est $\delta(t_1) \sqcap \delta(t_2) = (lcs(C(s_1), C(s_2)), lcs(p_1, p_2), lcs(C(o_1), C(o_2)))$ où lcs (*least common subsumer*) désigne la super-classe la plus spécifique ou le super-prédicat le plus spécifique, selon que les ressources soient des classes ou des prédicats respectivement.

Cette approche peut-être généralisée aux ensembles de triplets. La description d'un ensemble de triplets est l'ensemble des descriptions de chacun des triplets qu'il contient : $\Delta(T) = \{\delta(t) \mid t \in T\}$. La similarité entre deux descriptions d'ensembles de triplets $\Delta(X)$ et $\Delta(Y)$ est l'ensemble des similarités entre chaque paire de triplet $t_x \in \Delta(X)$ et $t_y \in \Delta(Y)$ les plus spécifiques. Une description (s_i, p_i, o_i) est dite *plus spécifique* qu'une autre description (s_j, p_j, o_j) si et seulement si s_i `subC` s_j, p_i `subP` p_j et o_i `subC` o_j.

3.2 Exemple

Pour calculer la similarité entre les deux ensembles de triplets, on commence par calculer la similarité paire à paire. Lorsque l'un des *lcs* renvoie un élément $\top$, il n'y a pas de similarité. Par exemple,

δ(`B_Obama, présidentDe, USA`)$\sqcap\delta$(`Arkansas, étatDe, USA`)
= (`Président, présidentDe, Pays`)$\sqcap$(`État, étatDe, Pays`)
= (lcs(`Président, État`), lcs(`présidentDe, étatDe`),
lcs(`Pays, Pays`))
= ~~($\top$, $\top$, `Pays`)~~.

Le triplet résultant est alors ignoré.

La similarité entre les deux ensembles de triplets Figure 1b est l'ensemble de cinq triplets suivant :

```
Président présidentDe Pays   Président néA Ville
Ville villeDe État           Lieu situéDans Lieu
État étatDe Pays
```

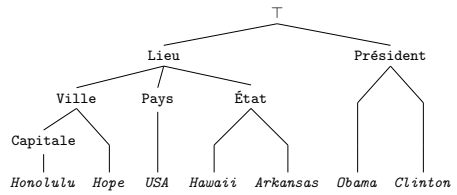
Ces cinq triplets représentent des connaissances qui, bien que triviales, n'étaient pas explicites. Ce sont ces éléments qui vont permettre de classifier des ensembles triplets.

4. CONCLUSION

L'approche proposée permet de classifier des ensembles de triplets en tenant compte des hiérarchies de classe et de prédicat. Elle a été testée sur des exemples simples et des tests sur des triplets issus de DBPedia sont en cours. Le travail à venir consiste en une implémentation de la structure de patrons proposée. Il s'agira également d'optimiser le procédé, en s'appuyant notamment sur RMQ [2]. Cette approche pourra ensuite étendre [1] pour extraire des définitions dans DBPedia.

5. REMERCIEMENTS

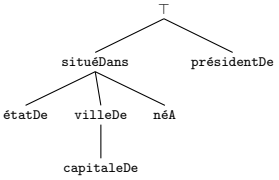
La thèse de Justine Reynaud est co-financée par la Direction Générale de l'Armement et par la Région Lorraine.



(a) Hiérarchie `subC` et instances associées

```
Ensemble 1
B_Obama présidentDe USA .
B_Obama néA Honolulu .
Honolulu capitaleDe Hawaii .
Hawaii étatDe USA .
Ensemble 2
B_Clinton présidentDe USA .
B_Clinton néA Hope .
Hope villeDe Arkansas .
Arkansas étatDe USA .
```

(b) ABox



(c) Hiérarchie `subP`

Figure 1: La figure 1b représente deux ensembles de faits (correspondant à la ABox des logiques de description). La figure 1c (1a) représente la hiérarchie de prédicats (classes) étant donné la relation `subP` (`subC`). Les instances sont reliées aux classes auxquelles elles appartiennent dans la figure 1a. Cet exemple est inspiré de [5].

6. REFERENCES

[1] M. Alam, A. Buzmakov, V. Codocedo, and A. Napoli. Mining Definitions from RDF Annotations Using Formal Concept Analysis. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, July 2015.

[2] M. Alam, A. Buzmakov, A. Napoli, and A. Sailanbayev. Revisiting Pattern Structures for Structured Attribute Sets. In *Proceedings of the 12th International Conference on Concept Lattices and Their Applications*, Clermont-Ferrand, France, Oct. 2015.

[3] M. Alam and A. Napoli. Interactive Exploration over RDF Data using Formal Concept Analysis. In *Proceedings of IEEE International Conference on Data Science and Advanced Analytics*, Paris, France, 2015.

[4] C. Carpineto and G. Romano. *Concept Data Analysis : Theory and Applications*. John Wiley & Sons, Chichester, UK, 2004.

[5] S. Ferré. A Proposal for Extending Formal Concept Analysis to Knowledge Graphs. In *Formal Concept Analysis*, volume LNCS 9113, pages 271–286, Nerja, Spain, June 2015.

[6] B. Ganter and S. O. Kuznetsov. Pattern structures and their projections. In *Conceptual Structures : Broadening the Base, 9th International Conference on Conceptual Structures*, Stanford, CA, USA, July 30-August 3.

[7] B. Ganter and R. Wille. *Formal concept analysis - mathematical foundations*. Springer, 1999.

[8] J. Kötters. Concept lattices of RDF graphs. *Formal Concept Analysis and Applications*, page 81, 2015.

[9] A. G. Nuzzolese, A. Gangemi, V. Presutti, F. Draicchio, A. Musetti, and P. Ciancarini. Tìpalo : A tool for automatic typing of dbpedia entities. In *The Semantic Web : ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers*, pages 253–257, 2013.

[10] B. Sertkaya. A survey on how description logic ontologies benefit from formal concept analysis. *CoRR*, abs/1107.2822, 2011.