Revealing the Conceptual Schemas of RDF Datasets - Extended Abstract

Subhi Issa, Pierre-Henri Paris, Fayçal Hamdi, Samira Si-said Cherfi

CEDRIC - Conservatoire National des Arts et Métiers
292 Rue Saint Martin, Paris, France

{subhi.issa,faycal.hamdi,samira.cherfi}@cnam.fr,pierre-henri.paris@upmc.fr

ABSTRACT. *This paper is an extended abstract of our work published at CAISE'20. The full paper is available at https://doi.org/10.1007/978-3-030-21290-2_20.*

RDF-based datasets, thanks to their semantic richness, variety and fine granularity, are increasingly used by both researchers and business communities. However, these datasets suffer a lack of completeness as the content evolves continuously and data contributors are loosely constrained by the vocabularies and schemes related to the data sources. In the context of the Web of Data and user-generated content, the conceptual schema is implicit. In fact, each data contributor has an implicit personal model that is not known by the other contributors. Consequently, revealing a meaningful conceptual schema is a challenging task that should take into account the data and the intended usage. In this paper, we propose a completeness-based approach for revealing conceptual schemas of RDF data. We combine quality evaluation and data mining approaches to find a conceptual schema for a dataset, this model meets user expectations regarding data completeness constraints. To achieve that, we propose LOD-CM; a web-based completeness demonstrator for linked datasets.

RÉSUMÉ. *Grâce à leur richesse sémantique, leur variété et leur granularité fine, les jeux de données fondés sur RDF sont de plus en plus utilisés par les chercheurs et les organisations. Cependant, ces jeux de données souffrent d'un manque de complétude en raison de l'évolution continue du contenu et le fait que les contributeurs ne sont pas tenus à respecter un vocabulaire et un schéma précis lors de la publication de leurs données. Dans cet article, nous proposons une approche fondée sur la complétude pour révéler les schémas conceptuels des données RDF. Nous combinons des approches d'évaluation de la qualité et de fouille de données pour trouver un schéma conceptuel pour un jeu de données, ce modèle répond aux attentes des utilisateurs en termes de complétude des données. Pour ce faire, nous proposons LOD-CM; un démonstrateur de complétude pour les jeux de données liés.*

KEYWORDS: *conceptual modeling, completeness, model quality, conceptual schema mining*

MOTS-CLÉS : *modélisation conceptuelle, complétude, qualité du modèle, extraction des schémas conceptuels*

1. Introduction

Data became a strategic asset in the information-driven world. One of the challenges for companies and researchers is to improve the display and understandability of the data they manage and use. However, exploiting and using data, like Linked Open Data (LOD), even if it is more and more accessible, is not an easy task, because data is often incomplete and lacks metadata. In this work we propose an approach for deriving conceptual schemas from existing data. This approach takes into account two facets; the universe of discourse represented by the data sources, and the user's needs represented by the user's decisions during the conceptual model construction. As the model should express the meaningful state of the considered dataset, we rely on a mining approach leading to taking into consideration the data model from a more frequent combination of properties. The relevancy of these properties is handled by integrating a completeness measurement solution that drives the identification of relevant properties. To meet user's requirements, we propose to construct the conceptual model on a *scratch card* manner where the user decides about the parts of the conceptual model to reveal according to her needs and constraints (Issa *et al.*, 2019).

2. Conceptual schemas derivation

The approach that we propose is an iterative process which infers a conceptual model complying the expected completeness. The process of inferring this model goes through four steps (cf. Figure 1): First, a subset of data that corresponds to the user's scope is extracted from the triple store. This subset is then transformed into transactions and a mining algorithm is applied. In our approach, for efficiency reasons, we chose the well-known FP-growth algorithm (Han *et al.*, 2004) (any other itemset mining algorithm could obviously be used). From the generated frequent itemsets, only a subset of these frequent itemsets, called "Maximal" (Grahne, Zhu, 2003), is captured. This choice is motivated by the fact that, on the one hand, we are interested in the *expression* of the frequent pattern and, on the other hand, the number of frequent patterns could be exponential when the transaction vector is very large. $\mathcal{MFP}$ is the set containing all maximal frequent patterns. Each pattern in $\mathcal{MFP}$ is then used to calculate the completeness of each transaction (regarding the presence or absence of the pattern) and, hence, the completeness of the whole dataset regarding this pattern. The final completeness value will be the average of all completeness value calculated for each $\mathcal{MFP}$ pattern. Finally, based on the completeness value and $\mathcal{MFP}$ that guarantees this value, a conceptual schema is generated. The classes, the attributes, and the relations of the model will be tagged with the completeness value. All these steps are integrated in an iterative process in such a way that the user could choose some parts in the generated model to refine. The data corresponding to the parts to refine is then extracted from the triple store, and the same steps are carried out to generate a new model.

Our prototype, called LOD-CM, implementing this process is available here: http://cedric.cnam.fr/lod-cm/.

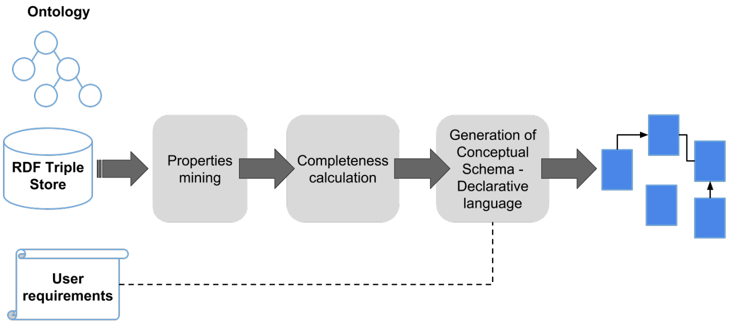

Figure 1. The LOD-CM Workflow.

3. Conclusion

In this paper, we presented an approach for revealing conceptual schemas from RDF data sources. Our approach is an iterative process that computes a plausible model from the data values. The inferred model takes into account the data and the user quality expectations. The result is a conceptual schema enriched by both completeness values as a relevancy indicator on the elements of the models, and existence constraints that inform about how often these elements co-exist or co-appear in the real data. In the future, we plan to investigate the role of conceptual modeling in an integration context where the universe of discourse is not only one data source but an integrated system upon several Linked Open Data. We plan to make more datasets available and allow the user to easily compare two conceptual schemas side by side (from two datasets). We believe that the ability to compare two conceptual schemas of two datasets side by side can help to choose the one that is best suited for its use.

References

Grahne G., Zhu J. (2003). Efficiently using prefix-trees in mining frequent itemsets. In B. Goethals, M. J. Zaki (Eds.), *FIMI '03, frequent itemset mining implementations, proceedings of the ICDM 2003 workshop on frequent itemset mining implementations, 19 december 2003, melbourne, florida, USA*, Vol. 90. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-90/grahne.pdf

Han J., Pei J., Yin Y., Mao R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, Vol. 8, No. 1, pp. 53–87. Retrieved from https://doi.org/10.1023/B:DAMI.0000005258.31418.83

Issa S., Paris P., Hamdi F., Cherfi S. S. (2019). Revealing the conceptual schemas of RDF datasets. In P. Giorgini, B. Weber (Eds.), *Advanced information systems engineering - 31st international conference, caise 2019, rome, italy, june 3-7, 2019, proceedings*, Vol. 11483, pp. 312–327. Springer. Retrieved from https://doi.org/10.1007/978-3-030-21290-2_20