

Une approche d'évolution du web de données

Fatma Chamekh*, Danielle Boulanger*, Guilaine Talens*

*Magellan-IAE-Jean Moulin University
6, cours Albert Thomas.69008 Lyon, France
(fatma.chamekh, danielle.boulanger, guilaine.talens) @univ-lyon3.fr,

1 Introduction

Dans cet article nous proposons une approche d'évolution du web de données fondée sur un appariement entre les deux parties de la base RDF : l'ontologie et les données. L'objectif est de modifier les deux niveaux simultanément. A la différence des méthodes existantes dont les liens entre ontologies et données sont parfois négligés (Papastefanatos et Stavrakas, 2014). Cette démarche est composée de trois étapes : l'extraction des données, l'identification des entités similaires et les propositions de changements.

2 L'extraction des données

Dans notre approche, l'utilisateur annote les documents non structurés via l'outil Dbpedia Spotlight. Le résultat est une ressource $R(l,c,uri,cs)$ mentionnant le label de la ressource (l), le/les URI, la/les classe(s) de l'ontologie (c) et le contextuel score qui mesure la représentativité de la ressource par rapport au contexte. Nous ajoutons un coefficient de pertinence qui mesure le degré d'interconnexion de la ressource à d'autres bases de connaissance. Celui-ci est le rapport entre les propriétés liées à la ressource R et le sous-ensemble des propriétés représentant le domaine d'application.

3 L'identification des entités similaires

Suite à la première étape, nous disposons de triplets $\langle r, rdf : type, c \rangle$ à ajouter. Pour trouver les correspondances dans la base RDF existante, nous emploierons des mesures de similarités syntaxiques (Sims) et d'autres sémantiques (sim_{PW}) pour calculer les similarités dans les deux niveaux.

4 Les propositions de changements

Pour proposer des changements, nous prenons en considération les mesures de similarités générées dans l'étape précédente. Il s'agit de combiner les mesures de similarités syntaxiques

Une approche d'évolution du web de données

et sémantiques pour lier les données et leurs classes. Afin de réaliser cette tâche nous utilisons l'algorithme suivant :

Algorithme 1 : proposition des changements

Entrées : $sim_s(R, R')$, $sim_s(C, C')$, $sim_{PW}(C, C')$ et les résultats annotation $\langle R, C \rangle$

```

1 début
2   Pour chaque sous ensemble pour chaque  $\langle R, C \rangle$  faire
3     si  $sim_s(R, R') \geq 0.6$  et  $sim_s(C, C') \geq 0.6$  alors
4        $R, owl : Sameas, R'$  et  $C, owl : Sameas, C'$ 
5     sinon
6       si  $sim_s(R, R') \geq 0.6$  et  $sim_s(C, C') \leq 0.6$  alors
7          $R, owl : Differentfrom, R'$  et  $C, owl : Differentfrom, C'$ 
8       sinon
9         si  $sim_s(R, R') \geq 0.6$  et  $Sim_{PW}(C, C') = 0$  alors
10           $R, owl : Sameas, R'$  et  $C, rdf : Subclassof, thing$ 
11        sinon
12          si  $sim_s(R, R') \geq 0.6$  et  $Sim_{PW}(C, C') \geq 0.6$  alors
13             $R, owl : Sameas, R'$  et  $C, rdf : Subclassof, C'$ 

```

5 Conclusion

Dans cet article nous avons présenté une approche d'évolution du web de données en modifiant simultanément les deux niveaux (A-Box et T-Box). Nous nous sommes basées sur des mesures de similarités syntaxiques et d'autres sémantiques pour proposer des changements. Notre perspectives est l'extension de notre approches pour pouvoir traiter des labels composés de phrases.

Références

Papastefanatos, G. et Y. Stavarakas (2014). Diachronic linked data: Capturing the evolution of structured interrelated information on the web. *Ercim news* 52, 35–37.

Summary

Sharing knowledge and data coming from different sources is one of the biggest advantage of linked data. Keeping this knowledge graph up to date may take in account both ontology vocabularies and data since they should be consistent. Our general problem is to deal with web of data evolution in particular: We aim at modifying both levels : A-Box and T-Box.