Approche numérique pour l'invalidation de liens d'identité (owl:SameAs)

Dimitrios Christaras Papageorgiou, Nathalie Pernelle, Fatiha Saïs

[1] Laboratoire de Recherche en Informatique, Université Paris Sud, Orsay, France
dimitrios.christaras-papageorgiou@u-psud.fr

[2] nathalie.pernelle@lri.fr

[3] Fatiha.Sais@lri.fr

Résumé : Au cours des dernières années, grâce à la standardisation des technologies Web sémantique, nous connaissons une production de données sans précédent, publiées en ligne sous forme de données liées. Dans ce contexte, lorsqu'un lien typé est déclaré entre deux ressources distinctes faisant référence à la même entité du monde réel, l'utilisation du owl :sameAs est généralement prédominant. Toutefois, des travaux récents dans la communauté des données liées ont montré des problèmes dans l'utilisation des liens owl :sameAs. Les problèmes surviennent à la fois dans les cas où ces liens sont erronés ou lorsqu'ils traduisent un lien moins strict que la sémantique des liens owl:sameAs définie dans OWL. Dans ce travail, nous présentons une méthode d'invalidation numérique de liens d'identité s'appuyant sur un calcul de similarité et sur des axiomes de l'ontologie pour détecter des liens d'identité invalides. Nous présentons nos premiers résultats expérimentaux, obtenus sur un jeu de données de la compétition internationale OAEI.

Mots-clés : Liage de données, Liens d'identité, Invalidation de liens, Ontologies, Qualité des liens et des données, Web Sémantique.

1 Introduction

Aujourd'hui, nous connaissons une production sans précédent de données, publiées sur le Linked Open Data (LOD), appelé aussi Web des Données. Cela a conduit à la création d'un espace de données global contenant des milliards d'assertions (Bizer *et al.* (2009)). Dans cet espace global, les fournisseurs de données établissent des liens RDF entre des ressources, représentées par des URIs, qui se réfèrent à la même entité du monde réel. Ainsi, on enrichit les connaissances relatives à une ressource spécifique et, par conséquent, la connaissance globale dans le Web des Données. La plupart des liens RDF reliant des ressources provenant de sources de données différentes sont des liens d'identité RDF représentés par le prédicat *owl*:*sameAs* dont la sémantique est défini dans Dean *et al.* (2004). Malheureusement, de nombreux liens d'identité existants ne reflètent pas une véritable identité et leur présence peut conduire à inférer des informations erronées et même contradictoires. Halpin *et al.* (2010) a ainsi montré que 37% de 250 liens d'identité choisis aléatoirement ont été déclarés comme erronés par cinq juges. De même, Jaffri *et al.* (2008) ont évalué la qualité du résultat d'une méthode de liage appliquée aux données de DBLP et DBpedia en choisissant arbitrairement 49 noms parmi les 491 796 auteurs disponibles dans DBLP de 2006. Ils ont montré que 92 % de ces 49 auteurs avaient alors des publications incorrectement affiliées. Compte tenu du volume important des données, les liens d'identité sont en effet souvent générés par des méthodes automatiques, avec des taux de précision inférieurs à 100 %. Les erreurs peuvent être dûes à la variation du niveau de qualité des données (i.e., complétude, correction, fraîcheur, fiabilité, etc.) entre les différentes sources de données liées, où à la difficulté de définir des règles de liage valides quelles que soient les

sources et les entités considérées. Ce problème montre la nécessité de définir des approches permettant de s'assurer de la qualité des liens. Certaines approches ont effectué une validation de type crowdsourcing pour évaluer et corriger les liens (Halpin *et al.* (2010)). D'autres approches utilisent la sémantique du lien d'identité, les axiomes de l'ontologies ou encore des hypothèses sur les données pour détecter autmatiquement qu'un lien conduit à une base de connaissance incohérente (de Melo (2013); Papaleo *et al.* (2014)). de Melo (2013) propose ainsi d'utiliser l'hypothèse du nom unique et la transitivité des liens d'identités pour détecter des inconsistances et utilise un algorithme de relaxation de contraintes pour supprimer des liens erronés. Papaleo *et al.* (2014) utilisent certains axiomes de l'ontologie (fonctionnalité) et la complétude locale de certaines propriétés pour invalider certains liens. D'autres travaux utilisent ces axiomes dans un cadre argumentatif pour générer des explications qui peuvent aider les experts à corriger les faits erronés (Arioua *et al.* (2016)). Les résultats d'approches telles que (Papaleo *et al.* (2014); de Melo (2013)) montrent que la précision des outils de liage peut réellement être améliorée quand elles sont utilisées pour filtrer les résultats. Cependant, dans un cadre purement logique, un fait erroné suffit à rendre la base de connaissances incohérente et ne permet pas de distinguer les cas où différents faits peuvent laisser penser que le lien est erroné de ceux dus à la présence d'un seul fait qui apparaît comme contradictoire.

Dans cet article, nous proposons une approche d'invalidation numérique de liens d'identité fondée sur des axiomes de fonctionnalité et sur les hypothèses de complétude-locale qui peuvent être déclarés pour certaines propriétés. Ce travail est une extension de Papaleo *et al.* (2014), dans lequel les axiomes sont utilisés dans un cadre numérique où différentes mesures d'agrégation simples peuvent être utilisées. Une première expérimentation a été menée sur des données de la compétition OAEI.

2 Approche numérique pour l'invalidation de liens d'identité

Le problème de détection de liens d'identité invalides se pose lorsque l'on souhaite vérifier si un lien d'identité *owl*:*sameAs* est valide entre deux ressources x et y dans un graphe RDF décrivant x et y. Plus précisément, nous souhaitons associer à un lien *owl*:*sameAs* un degré de confiance basé sur la similarité des descriptions des deux ressources. Notre approche exploite un graphe contextuel à profondeur n dont le contenu est délimité par un ensemble de propriétés P déclarées dans l'ontologie comme fonctionnelles, inverses-fonctionnelles ou locales-complètes. Plus précisément, la notion de graphe contextuel à profondeur n peut être défini comme suit (Papaleo *et al.* (2014)) :

Définition (**Graphe RDF**). Soit un ensemble U d'URIs, un ensemble B de nœuds blancs et un ensemble L de littéraux, un triplet RDF $< s, p, o >$ est tel que le sujet $s \in (U \cup B)$, le prédicat $p \in U$ et l'objet $o \in (U \cup B \cup L)$. Un graphe RDF G est une collection de triplets RDF.

Définition (**Chemin de propriétés de longueur** n) Soient G un graphe RDF, s un nœud dans G, et étant donné P un ensemble de propriétés défini pour G, un chemin de propriété $w_{n,s,P}$ de longueur n est une séquence alternant des nœuds et des propriétés, initiée par le nœud représentant la ressource s, $\{v_0 \equiv s, p_0, v_1, p_1, v_2, \ldots, v_{n-1}, p_{n-1}, v_n\}$, telle que : $v_0, \ldots, v_{n-1}$ sont des ressources dans G, $\forall i = 0, \ldots, n-1$ $v_i \in U$, v_n est un littéral dans G, $v_n \in L$, chaque triplet $\{v_i, p_i, v_{i+1}\}$ est une séquence dans un graphe RDF G telle que $p_i \in P$, toutes les ressources d'un chemin sont deux-à-deux distinctes.

Ce chemin peut être vu comme une collection d'assertions sélectionnées pour une ressource de départ s et un ensemble de propriétés P.

Définition (**Graphe Contextuel $G_{\{m,s,P\}}$ à profondeur** m). Soient G un graphe RDF et $s \in U$, un nœud de G, un nombre entier m et un ensemble P de propriétés défini pour G, un graphe contextuel $G_{\{m,s,P\}}$ à profondeur m pour une ressource s est un sous-graphe de G tel que chaque nœud $v_i \in G_{\{m,s,P\}}$ appartient à un chemin de propriétés de longueur n, avec $n \leq m$.

Un graphe contextuel à degré m pour une ressource s peut être considéré comme un sous-ensemble des informations pertinentes pour s, délimité par l'ensemble de propriétés P.

Définition (**Similarité contextuelle entre deux ressources** $CSim_{\{P,m\}}(x,y)$). Soient $G_{\{m,x,P\}}$ et $G'\{m,y,P\}$ deux graphes contextuels à profondeur m pour x et y, avec $P = DP \cup OP$ le sous-ensemble des propriétés de type *owl*:*DatatypeProperty* (DP) et de type *owl*:*ObjectProperty* (OP) délimitant le contexte dans les deux graphes G et G'. La similarité contextuelle pour les deux ressources x et y peut être définie comme suit :

$$CSim_{\{P,m\}}(x,y) =$$

$$F\Big(\bigcup_{\forall p_i \in DP} Sim(p_i.value(x), p_i.value(y))) \bigcup_{\forall p_j \in OP} CSim_{\{P,m\}}(p_j.value(x), p_j.value(y))\Big)$$

où :

— $p_i.value(x)$ permet d'obtenir la valeur ou les valeurs (en cas de propriétés multi-valuées) d'une propriété p_i de $G_{\{m,x,P\}}$,

— $Sim(v_x, v_y)$ est une fonction qui calcule un score de similarité dans $[0..1]$ entre v_x et v_y. Il s'agit soit de mesures de similarité élémentaires (e.g. Jacard, Jaro, Lenvenstein), soit de mesures de similarité entre ensembles de valeurs dans le cas de propriétés multi-valuées,

— F est une fonction d'aggrégation telle que la moyenne ou le minimum.

Définition du problème de détection de liens d'identité invalides : Étant donné un graphe RDF G, deux ressources x et y du graphe G, le triplet $< x, owl{:}sameAs, y >$ appartenant à G, un ensemble de propriétés P de G, un nombre entier m représentant la profondeur du graphe contextuel, deux graphes contextuels $G_{\{m,x,P\}}$ pour x et $G'_{\{m,y,P\}}$ pour y, un seuil de similarité $T \in [0..1]$, le problème d'invalidation numérique revient à déterminer si pour un couple de ressources x et y, on a : $CSim_{\{P,m\}}(x,y) \leq T$.

Comme cela a déjà été montré dans l'approche logique d'invalidation de liens d'identité Papaleo *et al.* (2014) le choix du sous-ensemble de propriétés à considérer peut être guidé par certains axiomes de l'ontologie : les propriétés fonctionnelles et inverses fonctionnelles et les propriétés locales complètes. En effet, quand une propriété p_1 est fonctionnelle, sa sémantique logique peut être exprimée par : $p_1(r,v) \wedge p_1(r,v') \Rightarrow v = v'$. Aussi la présence de valeurs différentes peut participer à l'invalidation d'un lien. De plus, si l'hypothèse du monde clos est en général inappropriée pour le Web sémantique (Heflin & Muñoz-avila (2002)), dans certains domaines et contextes spécifiques, la complétude locale de certaines propriétés peut être garantie (Wagner (2003)). Un bon exemple de propriété locale complète peut être la liste des auteurs d'un article dans un jeu de données tel que DBLP. Pour utiliser le fait qu'une propriété p est

locale-complète dans le calcul de similarité, nous vérifions que les ensembles de valeurs sont identiques. Aussi, nous vérifions que les deux ensembles de valeurs sont de même taille (si ce n'est pas le cas le score de similarité est mis à zéro), et si c'est le cas nous agrégeons les scores de similarité des paires de valeurs mises en correspondances en utilisant la fonction d'agrégation F qui a été sélectionnée.

En Figure 1, nous montrons un exemple de graphes contextuels extraits pour chercher à invalider un lien d'identité entre deux livres $b1$ et $b2$. La profondeur $m = 2$ et l'ensemble de propriétés P a été défini comme *{titre, annéeEd, auteur, nom}*. La propriété *ref*, étant non fonctionnelle et non locale complète, n'est pas prise en compte dans les graphe contextuels. La propriété *auteur* est déclarée comme locale complète. Considérons un score de similarité de 1 pour les valeurs de propriétés *titre* et *éditeur*, un score de 0 pour *annéeEd* et un score de 0.5 pour la propriété *auteur* en utilisant la mesure de similarité Jaccard. La fonction $CSim(b1, b2)$ avec $F = Moyenne$ donnerait un degré de confiance de 0.5 (0 avec $F = Minimum$).

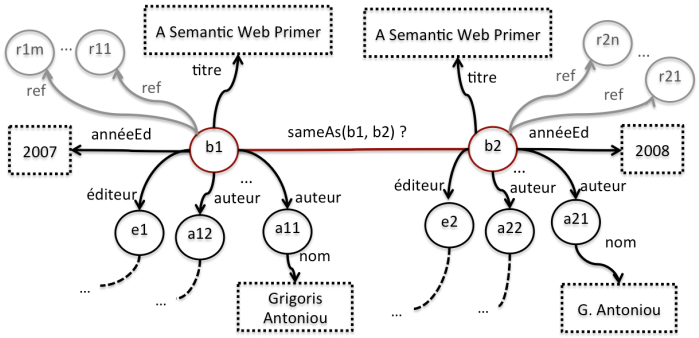

FIGURE 1 – Graphes contextuels extraits pour les ressources $b1$ et $b2$ (m=2)

3 Expérimentations

Nous avons évalué notre approche d'invalidation numérique de liens *owl :sameAs* sur les données du track PR de la compétition internationale OAEI [1] (Ontology Alignment Evaluation Initiative) 2010. Le benchmark *Person-Restaurants* a été conçu dans le but de mettre en compétition différents outils de liage de données. Les jeux de données *Person1* et *Person2* contiennent des données réelles et modifiées décrivant des personnes (SSN, nom, prénom, tél, et adresses décrites par des rues et villes), issues du projet Febrl [2]. Les données sont dupliquées de façon à ce que chaque description de *Person1* (resp. *Person2*) ait un doublon avec une modification au maximum (resp. trois modifications) par propriété. Le troisième jeu de données (*Restaurant*) a été créé en utilisant les descriptions de restaurants provenant de deux sources de données différentes. Les restaurants sont décrits par leur nom, adresse (rue, quartier et ville), téléphone et catégorie de restaurant (décrite par un nom de catégorie). Dans les trois jeux de données, le nombre d'instances varie entre 500 et 600. Pour chaque jeu de données, un goldstandard représentant le résultat de référence (i.e., un ensemble de 112 liens corrects) a été fourni. Pour tester

1. http://oaei.ontologymatching.org/2010/
2. http://sourceforge.net/projects/febrl/

notre approche, nous avons remplacé aléatoirement 50% des liens corrects par des liens erronés. Sur les trois jeux de données nous avons fait varier les mesures de similarité élémentaires (e.g.Jaro-Winkler, Jaccard), la fonction d'agrégation des scores de similarité ainsi que le seuil de similarité en dessous duquel un lien d'identité sera considéré comme invalide. La figure 2 montre les résultats en terme de rappel, précision et F-mesure obtenus sur les jeux (*Person1* et *Person2*). Nous avons comparé les résultats obtenus en utilisant la fonction moyenne (courbes en violet) et en utilisant la fonction minimum (courbes en jaunes) pour l'agrégation des scores de similarité. L'utilisation de la moyenne permet d'obtenir de bien meilleurs résultats que ceux obtenus avec un minimum, en terme de précision (et donc de F-Mesure). En effet, lorsque l'on utilise la fonction minimum, il suffit d'avoir une propriété pour laquelle le score de similarité est en dessous du seuil pour que cette similarité soit répercutée sur la similarité globale et conduise ainsi à une décision d'invalidation de lien (raisonnement analogue à celui de l'approche logique définie dans Papaleo *et al.* (2014)).

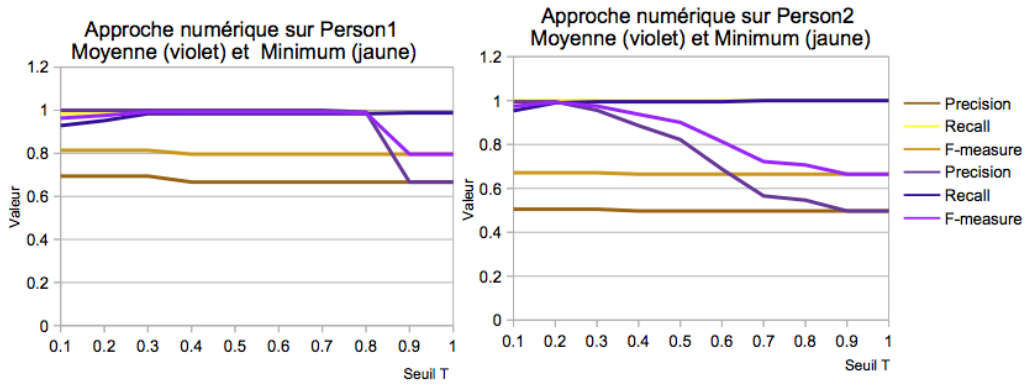

FIGURE 2 – Résultats (Précision, Rappel et F-mesure) de l 'approche numérique sur *Person1*

Les résultats obtenus sur *Restaurant*, sont similaires et montrent un gain maximum de F-Mesure de 20% quand la moyenne est utilisée. Les jeux sont très peu volumineux mais l'on peut noter que le temps d'exécution varie de 66 à 91 secondes pour les trois jeux de données (processeur Intel(R) Core(TM) i7-3630QM CPU@2.40GHz, mémoire RAM de 8Go).

Comparaison des résultats de l'approche numérique et de l'approche logique : Nous avons comparé nos résultats avec ceux obtenus par l'approche logique développée dans (Papaleo *et al.* (2014)). La Table 1 présente les résultats obtenus pour les deux approches sur les trois jeux de données *Person1*, *Person2* et *Restaurant*. Les résultats de l'approche numérique sont ceux obtenus au meilleur seuil de similarité et en utilisant la moyenne comme fonction d'agrégation des scores de similarité élémentaires. Sur les trois jeux de données les résultats de l'approche numérique en terme de F-mesure et précision sont meilleurs que les résultats de l'approche logique. En effet, nous obtenons un gain moyen de 23% de F-mesure en utilisant l'approche numérique, grâce à une augmentation très significative de la précision tout en ayant un résultat comparable en terme de rappel. Il suffit en effet d'avoir une seule propriété ayant des valeurs différentes pour que l'approche logique invalide le lien d'identité.

/	Approche logique (Papaleo *et al.* (2014))			Approche numérique			
Datasets	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure	Seuil
Person1	0.69	0.98	**0.81**	1.0	0.98	**0.99**	0.3
Person2	0.5	**1.0**	**0.67**	0.994	**0.989**	**0.99**	0.2
Restaurant	0.63	0.97	**0.77**	0.97	1.0	**0.98**	0.4

TABLE 1 – Comparaison entre l'approche logique Papaleo *et al.* (2014) et l'approche numérique

4 Conclusion

Nous avons présenté dans cet article une approche d'invalidation numérique de liens d'identité fondée sur le calcul d'un degré de confiance. Ce dernier exploite des sous-graphes RDF contextuels construits en prenant en compte des axiomes de (inverse) fonctionnalité des propriétés ainsi que des connaissances sur la complétude-locale de certaines propriétés. Les premières expérimentations ont montré la pertinence du choix d'une fonction d'agrégation simple comme la moyenne et le gain très significatif (de l'ordre de 23%) des résultats d'une telle approche par rapport à une approche purement logique (Papaleo *et al.* (2014)). Nous envisageons d'évaluer notre approche sur des jeux de données plus conséquents et de domaine différents. Nous souhaitons également combiner cette approche avec des approches de liage de données utilisant des règles efficaces de liage mais avec des exceptions.

Références

ARIOUA A., CROITORU M., PAPALEO L., PERNELLE N. & ROCHER S. (2016). On the explanation of sameas statements using argumentation. In *Scalable Uncertainty Management - 10th International Conference, SUM 2016, Nice, France, September 21-23, 2016, Proceedings*, p. 51–66.

BIZER C., HEATH T. & BERNERS-LEE T. (2009). Linked data - the story so far. *International Journal Semantic Web Information Systems*, **5**(3), 1–22.

DE MELO G. (2013). Not quite the same : Identity constraints for the web of linked data. In M. DESJARDINS & M. L. LITTMAN, Eds., *AAAI* : AAAI Press.

DEAN M., SCHREIBER G., BECHHOFER S., VAN HARMELEN F., HENDLER J., HORROCKS I., MCGUINNESS D. L., PATEL-SCHNEIDER P. F. & STEIN L. A. (2004). Owl web ontology language reference. *W3C Recommendation February*, **10**.

HALPIN H., HAYES P. J., MCCUSKER J. P., MCGUINNESS D. L. & THOMPSON H. S. (2010). When owl :sameas isn't the same : An analysis of identity in linked data. In *The Semantic Web – ISWC 2010 : 9th International Semantic Web Conference*, p. 305–320 : Springer Berlin Heidelberg.

HEFLIN J. & MUÑOZ-AVILA H. (2002). LCW-based agent planning for the semantic web. In *Ontologies and the Semantic Web Workshop*, p. 63–70 : AAAI Press.

JAFFRI A., GLASER H. & MILLARD I. (2008). Uri disambiguation in the context of linked data. In *Linked Data on the Web - LDOW*, volume 369 of *CEUR Workshop Proceedings* : CEUR-WS.org.

PAPALEO L., PERNELLE N., SAÏS F. & DUMONT C. (2014). Logical detection of invalid sameas statements in RDF data. In *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, p. 373–384.

WAGNER G. (2003). Web rules need two kinds of negation. In *Principles and Practice of Semantic Web Reasoning*, volume 2901 of *LNCS*, p. 33–50. Springer Berlin Heidelberg.