Fusion de données redondantes : une approche explicative

Fatiha Saïs[(1)] and Rallou Thomopoulos[(2)]

[(1)] LRI (CNRS UMR8623 & Université Paris Sud), Université Paris-Saclay,
Bât. 650 Ada Lovelace, F-91405 Orsay Cedex
[(2)] INRA (UMR IATE) & INRIA GraphIK, 2 place Viala, F-34060 Montpellier cedex 1

Résumé. Nous nous intéressons, dans le cadre du projet ANR Qualinca au traitement des données redondantes. Nous supposons dans cet article que cette redondance a déjà été établie par une étape préalable de liage de données. La question abordée est la suivante : comment proposer une représentation unique en fusionnant les "duplicats" identifiés ? Plus spécifiquement, comment décider, pour chaque propriété de la donnée considérée, quelle valeur choisir parmi celles figurant dans les "duplicats" à fusionner ? Quelle méthode adopter dans le but de pouvoir, par la suite, retracer et expliquer le résultat obtenu de façon transparente et compréhensible par l'utilisateur ? Nous nous appuyons pour cela sur une approche de décision multicritère et d'argumentation.

1 Introduction

Le *liage de données*, aussi appelé dans la littérature *réconciliation de données* (Ferrara et al. (2013); Saïs et al. (2009)), est le problème où l'on s'intéresse à détecter les descriptions référant au même objet du monde réel (e.g. même personne, même livre). La *fusion de données* est le problème posé suite à un liage de données si l'on souhaite fournir à l'utilisateur une représentation unique et homogène des données liées. La difficulté majeure rencontrée est celle des conflits et des inconsistances entre les valeurs d'une même propriété des données liées. Pour obtenir des données cohérentes et de bonne qualité, il faut résoudre ces conflits et choisir pour chaque propriété une ou plusieurs valeurs (cas des propriétés multi-valuées).

Des travaux ont étudié le problème de fusion de données dans le domaine des bases de données relationnelles (voir Bleiholder et Naumann (2009) pour un état de l'art). Dans cet article, nous nous intéressons au contexte du Web de données et étudions le problème de fusion de données RDF, distinct du cadre relationnel car caractérisé par la souplesse intrinsèque au modèle permettant la multi-valuation des propriétés, l'hypothèse du monde ouvert et la possibilité d'avoir plusieurs ontologies (schémas) décrivant les données. Récemment, des travaux (Saïs et Thomopoulos (2008); Flouris et al. (2012); Mendes et al. (2012)) se sont intéressés au problème de fusion de données RDF. Cependant aucun ne permet une bonne compréhension par l'utilisateur des résultats de la fusion. C'est sur cet aspect explicatif, visant à tracer et restituer les raisons ayant conduit à un résultat de fusion, que nous nous focalisons dans cet article.

La partie 2 décrit le problème de fusion et l'exemple illustratif. La partie 3 présente la méthode multicritère de fusion de données. La partie 4 introduit le processus explicatif qui s'appuie sur la construction d'arguments. La partie 5 conclut cette étude.

2 Problème de fusion et exemple illustratif

2.1 Problème de fusion de données

Nous considérons un ensemble d'instances $I = \{i_1, i_2, \ldots, i_n\}$ qui sont deux à deux liées par un lien d'identité (*owl :sameAs*). Chacune de ces instances est décrite par un ensemble de propriétés $P = \{p_1, p_2, \ldots, p_m\}$ déclarées dans une ontologie commune ou plusieurs ontologies alignées. Dans la suite de l'article, l'ensemble I sera désigné par *classe d'équivalence* d'instances et l'union des valeurs d'une propriété $p \in P$ pour toutes les instances $i \in I$ est appelé *ensemble de valeurs possibles*.

L'objectif de la fusion de données est d'obtenir une représentation unique de l'ensemble d'instances I de façon à pouvoir fournir la meilleure valeur ou le meilleur sous-ensemble minimal de valeurs (dans le cas de propriétés multi-valuées) pour chaque propriété $p \in P$.

2.2 Exemple illustratif

Tout d'abord, considérons un ensemble de données représentant trois descriptions différentes d'un même livre $b1, b2$ et $b3$. Chaque livre est décrit par l'ensemble de propriétés suivant (description tirées de `schema.org`) : {*title, nbPages, author, contributor, publisher, dateCreated, dateModified, datePublished, keywords*}.

Nous supposons qu'un outil de liage de données a été préalablement appliqué et a renvoyé le résultat suivant :

$<$ b1 $>$ owl : sameAs $<$ b2 $>$. $<$ b1 $>$ owl : sameAs $<$ b3 $>$. $<$ b2 $>$ owl : sameAs $<$ b3 $>$.

@prefix ob : $<$ https : //schema.org/Book $>$.

uri	ob :title	ob : nbPages	ob :author	ob :contributor	ob :publisher
b1	A Semantic Web Primer	238	Grigoris Antoniou	Paul Groth Frank V. Harmelen Rinke Hoekstra	The MIT Press (MA)
b2	A Semantic Web Primer	0	G. Antoniou	P. Groth F. V. Harmelen R. Hoekstra	MIT Press MA
b3	A Semantic Web Primer, second edition (cooperative information Systems Series	288	Grigoris Antoniou	Paul Groth Frank Van Harmelen Rinke Hoekstra	MIT Press Massachusetts

uri	ob :dateCreated	ob :dateModified	ob :datePublished	ob :keywords
b1	12/07/2007	01/05/2008	03/01/2008	Computer Science Knowledge representation Semantic Web
b2	December 7th 2007	April 30th 2004	March 1st 2008	Artificial Intelligence Description Logic Semantic Web
b3	December 2007	January 2008	March 2008	Semantic Web AI Knowledge representation & reasoning

FIG. 1 – *Exemple de trois descriptions d'un même livre*

En Figure 2 nous présentons un ensemble de règles d'incompatibilité entre valeurs de propriétés et de règles d'implausibilité de valeurs de certaines propriétés.

Les règles d'implausibilité :	R1 : nbPages ≤ 0
Les règles d'incompatibilité :	R2 : dateModified $\geq$ datePublished
	R3 : dateModified $\leq$ dateCreated

FIG. 2 – *Exemples de règles d'implausibilité et d'incompatibilité*

3 Méthode de décision pour la fusion de données

3.1 Le problème multicritère

Le choix de la valeur à retenir pour une propriété p de la donnée fusionnée peut être vu comme un problème de décision multicritère. Les deux entrées du problème, à savoir l'ensemble des options considérées d'une part, et l'ensemble des critères pris en compte pour les discriminer d'autre part, sont les suivants :

1. l'ensemble des options considérées est l'ensemble des valeurs prises par p dans I ;
2. l'ensemble des critères considérés comprend : la plausibilité, la précision, la synonymie, la compatibilité, l'homogénéité, la fréquence, la fraîcheur et la fiabilité de la source.

La section suivante fournit des précisions sur cet ensemble de critères et leur utilisation.

3.2 Méthodologie de fusion multicritère

Dans Giannopoulou et al. (2015), les auteurs ont développé une approche de fusion qui s'appuie un ensemble de critères et procède selon les étapes suivantes :

Prétraitement des valeurs peu plausibles. Une étape de pré-traitement des valeurs détecte les valeurs *peu plausibles* conformément à certaines contraintes de domaine et de typage de propriétés pour ainsi les exclure de l'ensemble de valeurs possibles de la propriété. Par exemple, si la propriété *pages* est typée comme "*xsd :nonNegativeInteger*", alors une valeur négative *pages* devrait être considérée comme peu plausible.
Découverte de relations de précision. Cette étape permet de découvrir des relations de type "*plus-précis-que*" entre les valeurs d'une même propriété dans la classe d'équivalence en utilisant : (i) des comparaisons syntaxiques en chaînes de caractères ou (ii) en exploitant des relations de subsumption ou de méréologie (*partie-de*) que l'on peut trouver dans des classifications de concepts et de termes. Par exemple, la valeur *"Knoweldge Representation"* est plus précise que *"Computer Science"* ; *"Massachusetts"* est plus précis que *"USA"*.
Découverte de relations de synonymie. Cette étape permet de découvrir des relations de synonymie entre les valeurs d'une même propriété dans la classe d'équivalence en utilisant des dictionnaires ou le résultat d'outils automatiques de liage de données. Par exemple, la valeur *"AI"* est synonyme de *"Artificial Intelligence"*.
Découverte de relations d'*incompatibilité.* Cette étape permet d'identifier les valeurs violant des règles de compatibilité, fournies à dire d'expert, pour une ou plusieurs propriétés.
Calcul d'un score de qualité pour chaque valeur. Dans cette étape, un score de qualité est calculé pour toute valeur jugée comme plausible. Ce score est calculé en fonction de critères

liés à la valeur elle-même (homogénéité et fréquence) et de critères liés à la source de données (fraîcheur et fiabilité). Pour plus de détails, voir Saïs et Thomopoulos (2008).

4 Explication des décisions de fusion

4.1 L'argumentation pour expliquer une décision

L'argumentation (Dung, 1995) s'avère un outil pertinent lorsque les avantages et les inconvénients doivent être évalués sur la base des connaissances disponibles. Très peu d'études abordent l'intérêt de l'argumentation pour la décision, les deux méthodes ayant historiquement été étudiées séparément avec des objectifs différents. Amgoud et Prade (2009) envisagent un processus en deux temps : 1) évaluation de l'ensemble des arguments construits pour ou contre les différentes options, 2) à partir des arguments acceptés lors de l'étape 1, classement des options par le choix d'une méthode d'agrégation des préférences. Notre proposition a la particularité d'utiliser les arguments comme "outils qualité", pour la traçabilité d'une décision. Nous introduisons pour cela la définition suivante d'un système de décision explicatif.

Définition *Un **système de décision explicatif** est un tuple $\langle D, G, A\rangle$ où :*

- *D est un ensemble d'options ;*
- *G est un ensemble de critères. A tout critère g est associé un domaine de valeurs $V(g)$;*
- *A est un ensemble d'arguments.*
  Tout $a \in A$ est défini par un quadruplet $\langle d_a, type_a, g_a, v_a\rangle$, où :
  - *$d_a \in D$ est l'option considérée par l'argument a ;*
  - *$type_a$ est un booléen indiquant le type de l'argument a, en faveur $(true)$ ou en défaveur $(false)$ de l'option d_a ;*
  - *$g_a \in G$ est le critère sur lequel s'appuie l'argument a pour s'exprimer en faveur ou en défaveur de l'option d_a ;*
  - *$v_a \in V(g_a)$ est la valeur prise par g_a dans l'option d_a.*

A partir du système de décision explicatif ci-dessus, deux fonctions permettent de fournir l'**explication du choix ou de l'élimination** d'une option :

$\forall d \in D$, $F_+(d) = \{a \in A | type_a = true \ \wedge \ d_a = d\}$ associe à chaque option d l'ensemble des arguments en faveur de d ;

$\forall d \in D$, $F_-(d) = \{a \in A | type_a = false \ \wedge \ d_a = d\}$ associe à chaque option d l'ensemble des arguments en défaveur de d.

4.2 Construction des arguments : processus et exemples

Reprenons pas à pas les étapes de la méthode de fusion indiquées dans la partie 3.2. Nous nous plaçons dans le cadre où une seule propriété p est considérée à la fois.

Valeurs peu plausibles. En cas de détection d'une valeur v peu plausible pour p, un argument en défaveur de v (qui est ici une option) est construit. Le critère sur lequel s'appuie cet argument est la plausibilité, qui a une mauvaise évaluation définie par les contraintes du domaine.

Exemple 1 *D'après la règle d'implausibilité $R1$ donnée en figure 2, dans les données de la figure 1 la valeur 0 pour la propriété* nbPages *est peu plausible. Cette valeur est rejetée et l'argument a_1 suivant est construit en défaveur de cette valeur : $a_1 = \langle 0, false, plausibility, \neg R1\rangle$.*

Relations de précision. En cas de détection d'une relation de précision entre deux valeurs v et

v' de la propriété p, deux arguments sont construits, l'un en faveur de l'option la plus précise, considérée comme étant la plus informative, l'autre en défaveur de l'option la moins précise. Le critère sur lequel se fondent ces arguments est la précision.

Exemple 2 *Pour la propriété* dateCreated *de la figure 1, la valeur "December 7th 2007" est identifiée comme plus précise que "December 2007". Les arguments suivants sont construits :*
$a_2 = \langle \text{“}December\ 7th\ 2007'', true, precision, > \text{“}December\ 2007'' \rangle$.
$a_3 = \langle \text{“}December\ 2007'', false, precision, < \text{“}December\ 7th\ 2007'' \rangle$.

Relations de synonymie. La détection d'une relation de synonymie ne permet pas à elle seule d'éliminer ou de sélectionner l'une ou l'autre des valeurs synonymes. Elle nécessite des informations supplémentaires, soit permettant d'opter pour un des synonymes en fonction de sa pertinence, soit autorisant des données multivaluées pour la propriété p, ce qui justifierait de conserver les 2 synonymes. Toutefois la détection de la synonymie accrédite les 2 valeurs, en établissant qu'elles ne sont pas aberrantes puisqu'il s'agit de variantes ayant la même signification. Pour cette raison, deux arguments sont construits, en faveur des deux options.

Exemple 3 *Pour la propriété* dateCreated *de la figure 1, les valeurs "December 7th 2007" et "12/07/2007" sont identifiées comme synonymes. Les arguments suivants sont construits :*
$a_4 = \langle \text{“}December\ 7th\ 2007'', true, synonymy, = 12/07/2007 \rangle$.
$a_5 = \langle 12/07/2007, true, synonymy, = \text{“}December\ 7th\ 2007'' \rangle$.

Relations d'incompatibilité. La détection d'une incompatibilité entre deux valeurs v_1 et v_2, pour deux propriétés distinctes p_1 et $p2$, introduit un doute sur ces deux valeurs. De ce fait, deux arguments sont générés, en défaveur des deux options. Toutefois, le contexte et les autres arguments en faveur ou en défaveur de v_1 et v_2 peuvent permettre d'identifier laquelle des 2 valeurs est erronée, ou si les deux le sont.

Exemple 4 *Dans la description $b2$ (Figure 1), les valeurs des propriétés* dateCreated *et* dateModified *violent la règle $R3$ (Figure 2). Deux arguments sont générés.*
Pour la propriété dateCreated *:* $a_6 = \langle \text{“}December\ 7th\ 2007'', false, compatibility, \neg R3 \rangle$.
Pour la propriété dateModified *:* $a_7 = \langle \text{“}April 30th 2004'', false, compatibility, \neg R3 \rangle$.
Un examen plus approfondi (autres arguments en faveur ou défaveur de chaque valeur, parmi lesquels a_2 et a_4) permet par la suite de conclure que la valeur erronée est celle de la propriété dateModified *("April 30th 2004") et d'éliminer l'argument a_6.*

Score de qualité. A ce stade, les arguments construits lors des étapes précédentes peuvent permettre de conclure sur la valeur à choisir pour la propriété p dans la donnée fusionnée. C'est par exemple le cas, dans la figure 1, des propriétés *author* et *contributor* en vertu d'arguments sur la précision, et de la propriété *dateModified* en vertu d'arguments sur la compatibilité et la précision. Le score de qualité, calculé pour toute valeur plausible, permet de classer les valeurs et de poursuivre le processus de fusion pour les propriétés restantes. Le calcul de ce score, qui n'est pas détaillé ici, s'accompagne également de la construction d'arguments, ainsi que de sous-arguments (pour chaque critère intervenant dans le score).

5 Conclusion

Cet article a introduit une méthode explicative de fusion, permettant à l'utilisateur de comprendre les résultats obtenus. Nous avons introduit un cadre de décision explicative, outil-qualité visant à tracer et restituer les raisons ayant conduit à la sélection ou à l'élimination d'une valeur dans la donnée fusionnée. Nous nous sommes appuyés pour cela sur une approche

combinant décision multicritère et argumentation. Les différents arguments *pour* et *contre* exploitent les connaissances utilisées lors du processus de prise de décision. Il s'agit notamment des relations de précision entre valeurs, du respect des contraintes de typage des propriétés, des connaissances sur la compatibilité entre valeurs de différentes propriétés, mais aussi d'informations qualitatives et quantitatives sur les valeurs (e.g. l'homogénéité, la fréquence) et sur leurs méta-données (e.g. la fraîcheur, la fiabilité des sources). Différentes directions sont à approfondir et en particulier le cas complexe des données multivaluées. Une autre perspective concerne l'ordre d'application des règles utilisées, et donc l'ordre de priorité des arguments construits. Cette approche générique n'est pas spécifique au cas de la fusion mais peut s'adapter et s'appliquer à d'autres contextes de décision.

Références

Amgoud, L. et H. Prade (2009). Using arguments for making and explaining decisions. *Artif. Intell. 173*(3-4), 413–436.

Bleiholder, J. et F. Naumann (2009). Data fusion. *ACM Comput. Surv. 41*(1), 1 :1–1 :41.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence 77*, 321–357.

Ferrara, A., A. Nikolov, et F. Scharffe (2013). Data linking. *J. Web Sem. 23*, 1.

Flouris, G., Y. R. andMaria Poveda-Villalon andPablo N. Mendes, et I. Fundulaki (2012). Using provenance for quality assessment and repair in linked open data. In *In Proceedings of the 2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn-12)*.

Giannopoulou, I., F. Saïs, et R. Thomopoulos (2015). Linked data annotation and fusion driven by data quality evaluation. In *15èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2015, 27-30 Janvier 2015, Luxembourg*, pp. 257–262.

Mendes, P. N., H. Mühleisen, et C. Bizer (2012). Sieve : linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pp. 116–123.

Saïs, F., N. Pernelle, et M. Rousset (2009). Combining a logical and a numerical method for data reconciliation. *J. Data Semantics 12*, 66–94.

Saïs, F. et R. Thomopoulos (2008). Reference fusion and flexible querying. In *Proceedings of On the Move to Meaningful Internet Systems : OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Part II*, pp. 1541–1549.

Summary

This paper deals with redundant data, previously identified by a data linkage step. The question considered is: how to propose a unique representation by merging the identified "duplicates"? More specifically, how to decide, for each data property, which value will be chosen among those describing the "duplicates"? What method should be adopted in order to be able to trace and explain the result in an understandable form to the user? The proposed approach relies both on multicriteria decision and argumentation.