Une méthode flexible de fusion de références

Fatiha Saïs*, Rallou Thomopoulos*,**

*LIRMM (CNRS & Univ. Montpellier II), 161 rue Ada, F-34392 Montpellier cedex 5
**INRA, UMR1208, 2 place P. Viala, F-34060 Montpellier cedex 1
Fatiha.Sais@lirmm.fr, rallou@supagro.inra.fr

Résumé. Dans cet article, nous nous intéressons au problème de fusion de références qui se pose une fois que les réconciliations entre références sont calculées. Il s'agit d'une tâche ayant comme objectif de fusionner les descriptions de références qui réfèrent à la même entité du monde réel pour en obtenir une seule représentation. Afin de pallier le problème d'incertitude dans les valeurs associées aux attributs, nous avons choisi de représenter les résultats de la fusion de références dans un formalisme fondé sur les ensembles flous. Nous indiquons comment les degrés de confiance sont calculés. Nous distinguons trois modes possibles de fusion. Enfin nous en proposons une représentation en RDF-Flou ainsi que son interrogation.

1 Introduction

La réconciliation et la fusion de données sont des problèmes majeurs pour l'intégration de données provenant de plusieurs sources. Ces problèmes sont liés à l'hétérogénéité syntaxique et sémantique du contenu des sources de données. La réconciliation consiste à décider si deux descriptions provenant de sources distinctes réfèrent ou non à la même entité du monde réel (e.g. la même personne, le même article, le même gène). La fusion consiste alors, à partir des descriptions réconciliées, à obtenir une seule représentation.

L'hétérogénéité des schémas est une des causes premières de la disparité de description des données entre sources (voir synthèse Rahm et Bernstein (2001); Shvaiko et Euzenat (2005)). Une autre cause d'hétérogénéité est due aux variations entre les descriptions des instances elles-mêmes. En effet, lors de l'intégration de données provenant de différentes sources, des représentations différentes peuvent référer la même entité du monde réel car des vocabulaires et des référentiels différents sont utilisés pour décrire les données. C'est dans ce cadre d'hétérogénéité liée aux données que nous nous situons dans cet article.

Comme les données sont créées de manière autonome et proviennent de différentes sources, nous ne pouvons pas faire l'hypothèse de l'identifiant unique. C'est pour cette raison que nous utilisons le terme *référence* d'une donnée au lieu d'*identifiant*. Nous parlerons alors des problèmes de réconciliation de références et de fusion de références.

Trouver les réconciliations entre références est néanmoins insuffisant pour obtenir une interrogation uniforme offrant à l'utilisateur des réponses non redondantes. Il est nécessaire d'avoir une méthode de fusion de descriptions des références réconciliées. C'est à ce problème que nous nous intéressons dans cet article. Les principales difficultés sont liées aux conflits

et aux ambiguïtés présentes dans les différentes descriptions de l'entité du monde réel. Différentes stratégies sont possibles pour choisir les valeurs à considérer dans la représentation finale de l'entité du monde réel. Dans ce travail, nous proposons une méthode flexible de fusion de réferences permettant plusieurs modes de fusion. Nous nous inspirons des travaux de Bleiholder et Naumann (2005), qui s'appuient sur des critères heuristiques ou statistiques comme par exemple la fréquence des valeurs, la fraîcheur des données ou encore la confiance dans les sources de données.

Par ailleurs, la certitude dans la résolution des conflits entre valeurs ne peut pas être garantie. Cette incertitude est d'autant plus présente lorsque les réconciliations entre références sont elles-mêmes incertaines. Afin de pallier ce problème, nous avons choisi de représenter les résultats de la fusion de références dans un formalisme fondé sur les ensembles flous (voir Zadeh (1965, 1978)). Nous proposons une méthode de calcul des degrés de confiance exploitant les caractéristiques des valeurs et des sources de données. Nous présentons l'utilisation d'une extension du langage RDF (RDF (2004)) pour la représentation de données floues. Nous montrons également comment le résultat de la fusion de références est représenté en RDF-Flou et interrogé par des requêtes SPARQL (SPARQL (2008)).

L'organisation de l'article est la suivante. La Section 2 présente la problématique de la fusion de références. La Section 3 donne nos contributions. La Section 4 en propose une implémentation en ce qui concerne la représentation des données fusionées

2 Problème de fusion de références

Nous considérons que nous disposons d'une méthode qui permet de décider de la réconciliation pour certaines paires de références. Nous partons donc d'un ensemble de réconciliations entre références pour lesquelles un score de similarité a été calculé.

En Figure 1 nous donnons un exemple de sources de données hétérogènes à réconcilier.

Source S1

Ref.	MuseumName	MuseumAddress	MuseumContact	PaintingName
id11	Lovre	Palais Royal, Paris	info@louvre.fr	La Joconde
id12	Louvre	Palais Royal, Paris	0140205317	Joconde
id13	Orsay	Rive gauche de la seine, Paris		L'Européenne

Source S2

Ref.	MuseumName	MuseumAddress	MuseumContact	PaintingName
id21	Louvre	99, rue Rivoli, 75001 Paris	info@louvre.fr 0140205317	Mona Lisa

FIG. 1 – *Exemple de références à fusionner*

En Figure 2 nous donnons un exemple d'ensemble de réconciliations, trouvées par la méthode N2R de réconciliation de références Saïs (2007), pour les références de la figure 1 .

A partir d'un ensemble de réconciliations entre références, une méthode de fusion de références a pour objectif de fournir, pour tout ensemble de références deux à deux réconciliées, une description de référence telle qu'il n'y ait pas de conflits dans les valeurs des attributs.

((id11, id21), 0,6) ; ((id11, id12), 0,9) ; ((id12, id21), 0,7)

FIG. 2 – *Ensemble de réconciliations entre les références*

Les différents cas de figure qu'une méthode doit considérer lors de la fusion de deux références réconciliées, sont les suivants : (i) un attribut dont la valeur est renseignée dans les descriptions de toutes les références ; (ii) un attribut dont la valeur est renseignée dans la description de certaines références et absente dans d'autres.

Comme nous sommes dans un contexte hétérogène où les variations syntaxiques sont fréquentes, une méthode de fusion ne peut pas garantir la certitude dans l'association des valeurs aux attributs de la référence résultant de la fusion. Par conséquent, au lieu d'avoir des couples $(attribut, valeur)$ issus de la fusion, on aura plutôt des triplets $(attribut, valeur, confiance)$ où $confiance$ est une valeur réelle dans $[0; 1]$ calculée lors de la fusion des références. Nous nous situons dans le cas où une référence d'une source peut être réconciliée avec plusieurs références, de la même source ou d'une autre source. Par conséquent, la fusion est appliquée sur un ensemble de références de l'ensemble des sources de données, deux - à - deux réconciliées.

3 Méthode flexible de fusion de références

Nous proposons une méthode pour la fusion de références qui permet de fournir un classement (ou *ranking*) des valeurs associées aux attributs. Ce faisant, cette méthode inclut la résolution des conflits entre valeurs telle que proposée dans des stratégies classiques déjà utilisées dans Bleiholder et Naumann (2005); Papakonstantinou et al. (1996). En effet, le classement des valeurs va permettre d'interroger les données fusionnées de manière flexible, en permettant de spécifier au niveau de l'interface d'interrogation le mode de fusion des références souhaité.

Nous apportons donc dans cette section deux contributions : une méthode de classement des valeurs prises par un attribut dans un ensemble de références réconciliées, par le calcul d'un degré de confiance, faisant l'objet de la partie 3.1 ; trois modes de fusion des références, correspondant à trois scénarios d'utilisation de cette méthode de classement, faisant l'objet de la partie 3.2.

3.1 Classement des valeurs d'un attribut

Objectif. Partant d'un ensemble de n références $ref_1, \ldots, ref_n$, telles que :

- chaque référence ref_i a pour description un ensemble de faits-attributs $Desc(ref_i) = \{< ref_i \ \ A_1 \ \ v_{i1} > \ldots < ref_i \ \ A_p \ \ v_{ip} >\}$;
- ces références sont deux à deux réconciliées, avec un score de similarité s_{ij} entre ref_i et ref_j.
- on note S_i la source de données dont est issue la référence ref_i ($S_1, \ldots, S_n$ non nécessairement distinctes),

l'objectif est de fournir, pour chacun des attributs A_k, la liste des valeurs v_{ik} prises par cet attribut, classées par un degré $c_{ik} \in [0; 1]$ mesurant la confiance dans le fait que la valeur correcte de l'attribut A_k soit v_{ik}.

Critères. Pour définir ce classement (ou ranking) sur les valeurs d'un attribut, nous considérons plusieurs critères et nous en proposons concrètement des mesures entre 0 et 1.

- en premier lieu, l'homogénéité des valeurs prises par l'attribut dans les références réconciliées. Nous proposons de mesurer cette homogénéité par la fréquence de chaque valeur parmi l'ensemble des valeurs prises par l'attribut dans les références réconciliées. Nous définissons donc l'homogénéité $hom(v_{ik})$ associée à la valeur v_{ik} comme suit.

$$hom(v_{ik}) = \frac{Card\{ref_j | < ref_j \ \ A_k \ \ v_{ik} > \in Desc(ref_j)\}}{n} \ \ avec\, j \ \in \ [1; n]$$

  Par exemple, dans le cas représenté en Figures 1 et 2, on a $hom(\text{“}Louvre'') = 2/3$;
- la similarité syntaxique des valeurs prises par l'attribut dans les références réconciliées. Ce critère repose sur l'hypothèse qu'une valeur est d'autant plus fiable qu'elle “ressemble” syntaxiquement aux valeurs prises par l'attribut dans les autres références considérées. Nous définissons la similarité $Csim(v_{ik})$ de la valeur v_{ik} avec les valeurs v_{jk} ($j \in [1; n]$, $j \neq i$) de la façon suivante.

$$Csim(v_{ik}) = \frac{\sum_j sim(v_{ik}, v_{jk})}{n-1}$$

  où sim désigne une mesure de similarité entre valeurs de base (voir Cohen et al. (2003)). Par exemple, dans le cas des Figures 1 et 2, on a $Csim(\text{“}Lovre'') = 5/6$, oú la valeur renvoyée par sim s'appuie sur le nombre de caractères communs entre deux chaînes.
- le score de similarité “global” entre les références réconciliées. Ce critère remplace le précédent lorsque la valeur de l'attribut n'est pas renseignée dans certaines références. On s'appuie alors sur la similarité de l'ensemble de la référence (i.e. prenant en compte les autres attributs) avec chacune des autres références réconciliées.
  Par exemple, dans le cas des Figures 1 et 2, si la valeur “Lovre” n'était pas renseignée, on aurait $Csim(null) = (0.6 + 0.9)/2 = 0.75$;
- la fraîcheur de la source de données. Ce critère est considéré comme une estimation de la fiabilité de la source de données.
  Nous proposons la définition suivante de la notion de fraîcheur. Soit $MAJ(S_i)$ la date de dernière mise à jour de la source de données S_i et j la date courante. La fraîcheur de S_i est donnée par :

$$frch(S_i) = 1 - \frac{j - MAJ(S_i)}{\sum_{p \in [1;n]}(j - MAJ(S_p))}.$$

  La forme de cette définition, qui fait intervenir le rapport entre l'ancienneté de mise à jour de la source considérée et la somme des anciennetés de mise à jour de toutes les sources, a l'intérêt suivant : elle tend vers la valeur 0 pour une source dont l'ancienneté de mise à jour est “écrasante” par rapport aux autres et vers la valeur 1 pour une source dont l'ancienneté de mise à jour est négligeable par rapport aux autres.
  Par exemple, si la dernière mise à jour de S1 a six mois et celle de S2 deux mois, on a $frch(S1) = 1 - 6/8 = 1/4$ et $frch(S2) = 1 - 2/8 = 3/4$;
- la fréquence d'occurrence des valeurs prises par l'attribut dans les références réconciliées, dans l'ensemble des valeurs de toutes les sources de données. En effet, une valeur

répétée à plusieurs reprises au sein des données sera considérée comme plus fiable, au sens où elle est moins susceptible de comporter des erreurs d'écriture (fautes de frappe, etc.) ou d'appartenir un à "jargon" propre à une source particulière.
La fréquence $f(v_{ik})$ de la valeur v_{ik} est définie par :

$$f(v_{ik}) = \frac{Card\{< ref \;\; A \;\; v_{ik} >\}}{\sum_{j\in[1;n]} Card\{< ref \;\; A \;\; v_{jk} >\}}$$

où ref désigne une référence appartenant à $S_1 \cup \ldots \cup S_n$ et A un attribut.
Par exemple, dans le cas des Figures 1 et 2, on a $f(\text{“}Lovre'') = 1/3$.

Détermination des degrés de confiance. Nous proposons une méthode s'appuyant sur les critères précédemment décrits pour associer à chaque valeur prise par un attribut dans l'ensemble de références réconciliées un degré de confiance entre 0 et 1.

Définition 1 *Soit A un attribut et $v_1, \ldots, v_n$ les valeurs respectives prises par A dans les références $ref_1, \ldots, ref_n$ deux à deux réconciliées. Le degré de confiance $conf(v)$, où $v \in \{v_1, \ldots, v_n\}$, mesurant la confiance dans le fait que la valeur correcte de l'attribut A soit v, est déterminé de la façon suivante :*

- *si $hom(v) = 1$ alors $conf(v) = 1$ (v est la valeur de A dans toutes les références) ;*
- *si $hom(v) < 1$ (v est la valeur de A dans certaines – mais pas la totalité – des références), soit I l'ensemble des indices $i \in [1; n]$ tels que $v_i = v$.*

Par exemple, dans le cas représenté en Figures 1 et 2, on a pour l'attribut *MuseumName* les valeurs "Lovre" et "Louvre", avec :
$hom(\text{“}Lovre'') = 1/3$ et $conf(\text{“}Lovre'') = (\frac{5}{6} + \frac{1}{4} + \frac{1}{3})/3 = 0.47$;
$hom(\text{“}Louvre'') = 2/3$ et $conf(\text{“}Louvre'') = max((\frac{11}{12} + \frac{1}{4} + \frac{2}{3})/3, (\frac{11}{12} + \frac{3}{4} + \frac{2}{3})/3) = max(0.61, 0.78) = 0.78$.

Nous obtenons ainsi un ensemble de valeurs possibles pour l'attribut A associées à un degré de confiance entre 0 et 1, soit la définition d'une distribution de possibilité (voir partie 4.1), définie sur $\{v_1, \ldots, v_n\}$, pour la valeur de l'attribut A.

3.2 Différents modes de fusion

Nous distinguons trois modes possibles de fusion. Le premier est la *fusion totale* des références, qui consiste à présenter, pour chaque attribut de la description d'une entité, la meilleure valeur, i.e. celle qui est au premier rang. Le deuxième est la *fusion partielle* des références, qui consiste à présenter, pour chaque attribut de la description d'une entité, les k premières valeurs (avec k une valeur entière positive). Ce mode de fusion peut s'appliquer dans le cas d'une interface d'interrogation, où le nombre de réponses à une requête est limité pour des raisons de lisibilité de l'interface graphique et de facilité d'utilisation. Le troisième est la *non fusion* des références, qui consiste à présenter de façon ordonnée toutes les valeurs associées aux attributs de la description d'une entité. Ce mode de fusion peut s'appliquer dans le cas d'un moteur de recherche d'informations, où le nombre de réponses potentielles n'est pas limité.

4 Implémentation de la fusion flexible de références

4.1 Rappels sur les ensembles flous

La théorie des ensembles flous a été introduite par Zadeh (1965) dans le but de formaliser la représentation de concepts vagues, tels que “jeune”, “proche” “rouge”, etc., qui ne peuvent pas être définis par des limites strictes.

Elle a donné naissance à la théorie des possibilités, également fondée par Zadeh (1978) et développée par Dubois et Prade (1988). Il s’agit d’une théorie de l’incertain.

Définition 2 *Une* distribution de possibilité π *est une fonction de* X *dans [0 ; 1] qui associe à chaque élément* x *de* X *le degré* $\pi(x)$ *avec lequel* $\{x\}$ *est possible.*

Une distribution de possibilité permet donc de représenter une donnée dont la valeur est incertaine.

Dans la suite de cet article, les données incertaines constituées par le résultat d’une fusion de références, seront définies par leurs distributions de possibilité, et représentées en RDF-Flou.

4.2 Représentation des données fusionnées en RDF-Flou

Afin de pouvoir représenter des données incertaines, Mazzieri (2004) a proposé d’étendre RDF en RDF-Flou en dféinissant une syntaxe et une sémantique.

L’extension de la *syntaxe* consiste à exprimer les déclarations RDF sous forme de triplets $< sujet, predicat, objet >$ par des couples de la forme $< valeur, triplet >$ en ajoutant donc une valeur réelle, dans $[0; 1]$, à chaque triplet. Il s’agit de préfixer chaque triplet par une valeur réelle, dans $[0; 1]$, représentant la valeur de vérité floue du triplet. Nous obtenons alors des déclarations RDF-Flou de la forme $n :< sujet\ predicat\ objet >$. Pour avoir plus de détails concernant cette extension de RDF au RDF-Flou, voir Mazzieri (2004).

Nous noterons ref_F la référence issue de la fusion, comportant des valeurs incertaines. En application de la Définition 1, nous obtenons en utilisant la syntaxe RDF-Flou une description de ref_F constituée d’un ensemble de déclarations de la forme :
$conf(v) :< ref_F\ A\ v >$
où A est un attribut et v une valeur parmi l’ensemble des valeurs distinctes prises par A dans les références fusionnées.

Ainsi l’exemple donné à la fin de la partie 3.1 fournit les déclarations suivantes, décrivant la distribution de possibilité de la valeur de l’attribut *MuseumName* :
$0.47 :< ref_F\ MuseumName\ “Lovre'' >$
$0.78 :< ref_F\ MuseumName\ “Louvre'' >$.

Pour assurer l’implémentation de notre méthode de fusion dans tout type de plate-forme basée sur le langage RDF, nous proposons une traduction de notre représentation RDF-Flou des données fusionnées en RDF. Pour ce faire, nous utilisons le mécanisme de réification dont la sémantique est définie dans Hayes (2004) et qui permet d’ajouter des éléments de description suplémentaires aux déclarations RDF, tels que l’auteur des données, la date de création etc.

Dans notre cas, la réification consiste à ajouter aux triplets de la forme $< ref_F\ A\ v >$ la propriété *confiance* dont le domaine est de type ressource et le co-domaine est de type décimal.

Avec cette transformation, les références fusionnées peuvent être interrogées en utilisant le langage SPARQL sans aucune extension.

4.3 Interrogation des données fusionnées

Afin d'illustrer l'interrogation des références fusionnées, nous allons utiliser des requêtes SPARQL de type *selection*. Ces requêtes sont évaluées sur les références fusionnées et représentées en RDF. La représentation RDF est obtenue grâce à la transformation du RDF-Flou en utilisant la réification.

La requête ci-dessous peut être adaptée au mode de fusion souhaité de la manière suivante :
- *fusion totale*, ne rien modifier ;
- *fusion partielle* en renvoyant les k-premières valeurs, il suffit de modifier la valeur suivant la clause LIMIT en remplaçant la valeur 1 par k ;
- *non fusion*, il suffit de supprimer la clause LIMIT et ainsi renvoyer toutes les valeurs ordonnées par le degré de confiance.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX monEspace: <http://www.lirmm.fr/~sais/myRDFS-1/>
SELECT  ?ref ?nom ?confiance
WHERE {
?x rdf:type rdf:Statement  .
?x rdf:subject ?ref  .
?x rdf:predicate monEspace:MuseumName .
?x rdf:object ?nom .
?x rdf:object ?confiance .
}
ORDER BY ?confiance
LIMIT 1
```

5 Conclusion

Dans cet article nous avons proposé une méthode de fusion de références qui permet de calculer pour chacune des valeurs intervant dans un conflit, un degré de confiance. Ce dernier est obtenu par la combinaison de plusieurs critères liés à la fois à la nature syntaxique des valeurs elles-mêmes mais aussi aux caractéristiques des sources de données. Les ensembles de valeurs associées aux attributs avec des degrés de confiance sont exprimés en utilisant les ensembles flous. Le résultat de cette méthode de fusion est réprésenté en RDF-Flou qui peut être trivialement transformé en RDF en utilisant la réification. Nous avons également montré comment on peut exploiter la flexibilité de cette approche lors de l'interrogation des références fusionnées. En effet, nous avons défini trois modes de fusion. Le choix du mode de fusion souhaité peut être effectué au niveau des requêtes.

Nous avons proposé une méthode locale qui ne traite que les conflits qui surgissent entre les valeurs des attributs. Il est cependant important de pouvoir également traiter des conflits qui peuvent apparaître entre les références liées à d'autres références par les relations. Par exemple, les références d'oeuvres contenues dans un musée. Nous envisageons également

d'étudier comment prendre en compte les préférences des utilisateurs lors de l'interrogation des données fusionnées et représentées par des ensembles flous.

Références

Bleiholder, J. et F. Naumann (2005). Declarative data fusion – Syntax, semantics, and implementation. In *Proc. of the 9th East European Conference on Advances in Databases and Information Systems*.

Cohen, W. W., P. Ravikumar, et S. E. Fienberg (2003). A comparison of string distance metrics for name-matching tasks. In *IIWeb*, pp. 73–78.

Dubois, D. et H. Prade (1988). An introduction to possibilistic and fuzzy logics. In P. Smets, A. Mamdani, D. Dubois, et H. Prade (Eds.), *Non-Standard Logics for Automated Reasoning*, pp. 287–315. London : Academic Press.

Hayes, P. (2004). RDF Semantics, http ://www.w3.org/tr/rdf-mt/. Technical report.

Mazzieri, M. (2004). A fuzzy rdf semantics to represent trust metadata. In *In 1st Workshop on Semantic Web. Applications and Perspectives*.

Papakonstantinou, Y., S. Abiteboul, et H. Garcia-Molina (1996). Object fusion in mediator systems. In *VLDB*, San Francisco, CA, USA, pp. 413–424.

Rahm, E. et P. A. Bernstein (2001). A survey of approaches to automatic schema matching. *The VLDB Journal 10*(4), 334–350.

RDF (2004). http ://www.w3.org/rdf/.

Saïs, F. (2007). *Intégration sémantique de données guidée par une ontologie*. Ph. D. thesis, université de Paris-Sud.

Shvaiko, P. et J. Euzenat (2005). A survey of schema-based matching approaches. pp. 146–171.

SPARQL (2008). http ://www.w3.org/tr/rdf-sparql-query/.

Zadeh, L. (1965). Fuzzy sets. *Information and Control 8*, 338–353.

Zadeh, L. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems 1*, 3–28.

Summary

This paper deals with the issue of data fusion, which arises once reconciliations between references have been determined. The objective of this task is to fusion the descriptions of references that refer to the same real world entity so as to obtain a unique representation. In order to deal with the problem of uncertainty in the values associated with the attributes, we have chosen to represent the results of the fusion of references in a formalism based on fuzzy sets. We indicate how the confidence degrees are computed.We distinguish between three possible modalities of fusion based on the fuzzy fusion result. Finally we propose a representation in Fuzzy RDF, as well as its querying.