Détection de redondances dans les tableaux guidée par une ontologie

Rania Khefifi*, Patrice Buche**,
Juliette Dibie-Barthélemy***, Fatiha Saïs*

* LRI/INRIA Saclay, 4 rue Jacques Monod, F-91893 Orsay Cedex, France
{Rania.Khefifi, Fatiha.Sais}@lri.fr
**INRA - UMR IATE, 2, place Pierre Viala, F-34060 Montpellier Cedex 2, France
LIRMM, CNRS-UM2, F-34392 Montpellier, France
Patrice.Buche@supagro.inra.fr
***INRA - Mét@risk & AgroParisTech, 16 rue Claude Bernard,
F-75231 Paris Cedex 5, France
Juliette.Dibie@agroparistech.fr

Résumé. Nous nous intéressons dans cet article à la réconciliation d'annotations floues associées à des tableaux de données par une méthode d'annotation sémantique, qui est guidée par une ontologie de domaine. Etant donnés deux tableaux, la méthode consiste à détecter leurs instances de relation redondantes. Elle s'appuie sur les connaissances déclarées dans l'ontologie, ainsi que sur des scores de similarité entre les annotations floues représentées par des sous-ensembles flous numériques ou par des sous-ensembles flous symboliques.

1 Introduction

L'ouverture sur le Web permet de publier des documents sans aucun contrôle sur leur contenu. Cet absence de contrôle a des avantages : la richesse et la diversité des informations disponibles sur le Web ; mais elle a également des inconvénients, notamment l'hétérogénéité et la redondance de ces informations. Dans le domaine de l'intégration de données, des travaux ont été menés sur la construction d'entrepôts thématiques de données extraites à partir de sources hétérogènes publiées sur le Web. Certains travaux, comme celui de Hignette et al. (2009), se sont en particulier focalisés sur l'extraction et l'intégration de données structurées représentées dans des tableaux. Les tableaux extraits proviennent de différentes sources (articles scientifiques, rapports de projets, mémoires de thèses, etc.) dans des formats hétérogènes (documents HTML, documents PDF ...). Leur intégration dans un entrepôt thématique repose sur une méthode d'annotation sémantique de tableaux, guidée par une ontologie de domaine, qui permet de traiter le problème de l'hétérogénéité sémantique du point de vue du vocabulaire utilisé pour décrire les données. Cette méthode génère automatiquement des annotations floues. Une annotation floue représente une instance de relation sémantique de l'ontologie reconnue sur une ligne d'un tableau. L'annotation sémantique n'empêche cependant pas l'intégration de données redondantes dans l'entrepôt. La présence de redondance dans l'entrepôt dégrade la

qualité de ses données et par conséquent leur exploitation (e.g. interrogation par l'utilisateur, analyse des données ou aide à la décision). Le problème de la détection de redondance est proche de celui de la réconciliation de références (Dong et al. (2005), Tejada et al. (2001), Bilenko et Mooney (2003)). Les méthodes supervisées de réconciliation de références supposent l'existence d'un ensemble de paires de références, qui sont étiquetées réconciliées ou non réconciliées. Ces méthodes utilisent des algorithmes d'apprentissage supervisé, qui exploitent cet ensemble d'exemples positifs et négatifs, pour apprendre des connaissances ou des paramètres nécessaires à la prise de décision de réconciliation ou de non réconciliation. Dans cet article, nous nous intéressons au problème de la détection de données redondantes dans des tableaux annotés par la méthode présentée dans Hignette et al. (2009), qui utilise une ontologie de domaine. Nous avons fait le choix de ne pas imposer une phase d'apprentissage à l'utilisateur de l'entrepôt pour résoudre ce problème. Nous proposons une approche déclarative, inspirée de la méthode N2R proposée par Saïs et al. (2009), dans laquelle des connaissances sur la disjonction entre classes et sur des propriétés sont déclarées dans l'ontologie. En section 2, nous présentons l'ontologie et la méthode d'annotation sémantique de tableaux de Hignette et al. (2009). Nous proposons ensuite en section 3 une nouvelle méthode de détection automatique de redondances dans les tableaux annotés.

2 Préliminaires

Nous présentons dans cette section l'ontologie de domaine utilisée dans notre méthode de détection de redondances. Nous décrivons ensuite brièvement la méthode qui est utilisée pour l'annotation sémantique de tableaux, étape préalable à la détection de redondances.

Présentation de l'ontologie. Il existe dans la littérature plusieurs définitions d'une ontologie (cf. Gruber (1993)). L'ontologie que nous considérons est un modèle de données représentant de manière structurée le vocabulaire d'un domaine spécifique sous la forme d'un ensemble d'attributs et de relations n-aires entre ces attributs. On distingue deux types d'attributs : les attributs symboliques et les attributs numériques qui sont représentés par les deux concepts *symbolic-attribute* et *numerical-attribute*, eux-mêmes subsumés par le concept *attribute*. Chaque attribut symbolique est défini par un nom et par une taxonomie de termes qui représente l'ensemble des termes possibles pour cet attribut (ces termes étant représentés par des concepts dans l'ontologie). Les attributs numériques sont définis par leur nom, leurs unités de mesure et, de manière optionnelle, par leur intervalle de valeurs possibles.

Une relation sémantique est définie par son nom et sa signature. La signature d'une relation sémantique est représentée par un ensemble de n ($n > 1$) attributs reliés entre eux. On distingue deux types d'attributs : les *attributs d'accès* qui représentent le domaine de la relation sémantique et un *attribut résultat* qui représente le co-domaine de la relation. Une relation sémantique est notée comme suit : $relSem(Aa_1, Aa_2, \ldots, Aa_n, Ar)$, où, $relSem$ représente le nom de la relation sémantique et $(Aa_1, Aa_2, \ldots, Aa_n, Ar)$ représente la signature de la relation sous la forme d'un ensemble d'attributs d'accès Aa_i et d'un attribut résultat Ar.

Exemple : La relation `ContaminationLevel(Food Product, Microorganism, Colony Count Concentration)` a comme ensemble d'attributs d'accès `{Food Product, Microorganism}` et comme attribut résultat `Colony Count Concentration`.

Méthode d'annotation. Un tableau est composé de colonnes, elles-mêmes composées de cellules. Dans Hignette et al. (2009), les cellules d'un tableau peuvent contenir des termes ou des valeurs numériques. L'annotation sémantique d'un tableau consiste à identifier les attributs symboliques ou numériques représentées par ses colonnes, puis identifier la ou les relations sémantiques n-aires qui existent entre ses colonnes. Enfin, les relations identifiées sont instanciées pour chaque ligne du tableau. Dans une instance de relation donnée, chaque valeur de cellule (associée à une instance d'attribut de la relation) est représentée par un sous-ensemble flou en fonction de la nature de la cellule : (i) les cellules numériques sont instanciées par des sous-ensembles flous à support numérique qui représentent une distribution imprécise de valeurs et (ii) les cellules symboliques sont instanciées par des sous-ensembles flous à support symbolique contenant chacun une liste de termes de l'ontologie les plus similaires à la valeur d'origine avec leur score de similarité.

3 Méthode de détection de redondances dans des tableaux

Nous présentons dans cette section notre méthode de détection de redondances. Elle permet de détecter les paires d'instances de relations redondantes dans des tableaux extraits du Web. La méthode prend en entrée deux tableaux annotés par un ensemble d'instances de relations sémantiques et renvoie un ensemble de paires d'instances de relations redondantes.

Définition 1 (instance de relation) : Une instance de relation sémantique ir est représentée par un couple $(id, desc_{ir})$, avec id l'identifiant de l'instance de relation et $desc_{ir}$ sa description. La description d'une instance de relation est composée de l'ensemble d'instances d'attributs appartenant à sa signature, $desc_{ir} = \{(a_1, inst_{a_1}), \ldots, (a_i, inst_{a_i}), \ldots, (a_n, inst_{a_n})\}$, avec $inst_{a_i}$ l'instance du i-ème attribut a_i de la relation sémantique (*c.f.* Définition 2).

Définition 2 (instance d'attribut) : Une instance $inst_{a_i}$ d'un attribut a_i est représentée par un sous-ensemble flou qui peut être :

1. numérique sous la forme d'un trapèze, avec un intervalle support $[S_{min}, S_{max}]$ et un intervalle noyau $[K_{min}, K_{max}]$ et est noté comme suit $inst_{a_i} = (a_i, [S_{min}, K_{min}, K_{max}, S_{max}])$.
2. symbolique sous la forme d'un ensemble de termes t_k de l'ontologie associés à leur degré d'appartenance d_k et est de la forme, $inst_{a_i} = (a_i, \{t_1/d_1, \ldots, t_k/d_k\})$.

Algorithme de détection de redondances. Étant donnés deux tableaux *T1* et *T2* annotés par une même ontologie de domaine, nous cherchons à détecter toutes les paires d'instances de relations redondantes. Pour ce faire, nous calculons un score de similarité pour chaque paire d'instances de relations. Nous supposons que nous avons en entrée de la méthode toutes les instances de relations sémantiques identifiées dans le tableau. Nous construisons d'abord l'ensemble des paires d'instances de relations comparables *IRC* en tenant compte des disjonctions entre relations sémantiques. Deux instances de relations (ir_1, ir_2) sont comparables si les relations sémantiques r_1 et r_2 ne sont pas déclarées disjointes dans l'ontologie, ce qui permet d'éviter des comparaisons inutiles. Nous calculons ensuite un score de similarité pour chaque paire d'instances de relations (ir_1, ir_2) de l'ensemble *IRC*. Finalement, une paire d'instances

(ir_1, ir_2) est dite redondante si son score de similarité est supérieur à un certain seuil prédéfini.

Calcul de similarité entre deux instances de relation. Le score de similarité d'une paire d'instances de relations sémantiques (ir_1, ir_2) est obtenu en combinant les scores de similarité des paires d'instances d'attributs qui décrivent les instances de relation (ir_1, ir_2). Le calcul de ce score est détaillé ci-dessous. Deux instances d'attributs sont comparables si elles instancient des attributs non déclarés disjoints dans l'ontologie. Pour chaque instance d'attribut de ir_1, un score de similarité est calculé avec toutes les instances d'attributs comparables de ir_2. Ce score est une combinaison des deux scores suivants : (1) $score_{Sem}$ qui est calculé à partir de la similarité sémantique entre les attributs si ces attributs sont différents ; cette similarité repose sur la notion du plus petit généralisant commun (LCS en anglais) entre les attributs dans la taxonomie de concepts et (2) $score_{Inst}$ qui est calculé à partir des annotations associées aux instances d'attributs. Nous utilisons deux mesures de similarité différentes selon que l'instance d'attribut est un sous-ensemble flou numérique ou symbolique. Ces mesures seront présentées dans les paragraphes suivants. Finalement, pour chaque instance d'attribut de la première relation, on retient le meilleur score de similarité avec les instances d'attributs de la deuxième relation. Lorsque toutes les paires d'instances d'attributs et leurs scores ont été calculés, nous pouvons calculer le score de similarité de la paire d'instances de relations défini comme le produit des scores des paires d'instances d'attributs.

Comparaison des annotations floues associées à deux instances d'attributs numériques. Nous proposons une nouvelle mesure inspirée de la méthode de centre de gravité simple MCGS proposée par Chen et Chen (2003). Elle repose sur un calcul de similarité entre les centroïdes des sous-ensembles flous à comparer. Nous commençons par calculer les coordonnées des centroïdes des deux sous-ensembles flous, notés x^* et y^*, comme suit (cf. formules 1 et 2). Soit un sous-ensemble flou à valeur numérique de forme trapezoïdale (a_1, a_2, a_3, a_4),

$$\text{Si } a_1 = a_4 \longrightarrow \begin{cases} y^* = 1/2 \\ x^* = a_1 \end{cases} \quad (1) \qquad \text{Sinon} \longrightarrow \begin{cases} y^* = \frac{\frac{a_3 - a_2}{a_4 - a_1} + 2}{6} \\ x^* = \frac{y^*(a_3 + a_2) + (a_4 + a_1)(1 - y^*)}{2} \end{cases} \quad (2)$$

Nous calculons ensuite la distance euclidienne $d(cent_A, cent_B)$ entre les deux centroïdes des deux sous-ensembles flous trapézoïdaux A et B. Nous pouvons enfin calculer la mesure de similarité : $score_{Inst}(A, B) = 1/(1 + d(cent_A, cent_B))$.

Comparaison des annotations floues associées à deux instances d'attributs symboliques. Nous proposons une nouvelle mesure, appelée $Jaccard_{flou}$, inspirée de la mesure de Jaccard (Jaccard (1901)) qui repose sur le rapport entre le nombres de termes en commun (somme des degrés minimaux des termes en communs) et le nombre total de termes (somme des degrés maximaux des termes) des deux ensembles flous. Soient A et B deux ensembles flous symboliques, $deg_A(t)$ (respectivement $deg_B(t)$) le degré d'appartenance du terme t dans l'ensemble A (respectivement B), nous avons : $score_{Inst}(A, B) = Jaccard_{flou}(A, B)$

$$Jaccard_{flou}(A, B) = \Big(\sum_{t \in A \cap B} \min(deg_A(t), deg_B(t))\Big) \,/\, \Big(\sum_{t \in A \cup B} \max(deg_A(t), deg_B(t))\Big)$$

Exemple de calcul : Soient les deux tableaux Tab. 1 et Tab. 2 extraits du web :

Les relations reconnues dans le 1^{er} tableau (resp. le 2^{eme} tableau) sont les suivantes : (i) *ContaminationLevelRelation (Food, Contaminant, year, ContaminationLevel)* ; *LodRelation (Food, Contaminant, year,*

Food	*Contaminant*	*Year*	*Lod*	*Contamination Level*
Baby food	Patulin	2000	0.7	6.3
Apple juice	Patulin	1998	2	8.37
Breafast cereal	Ochratoxin A	2003	0.7	<0.2

Tab. 1– Premier exemple de tableau du Web

Food	*Contaminant*	*Max Value*	*Contamination Level*
Baby food	Patulin	58	6.3
Breakfast cereals	Ochratoxin A	6	<0.2

Tab. 2 – Deuxième exemple de tableau du Web

lod) et (ii) *ContaminationLevelRelation (Food,Contaminant, year, ContaminationLevel)* ; *MaxContaminantionLevelRelation (Food, Contaminant, year, MaxLevel).*

Les attributs soulignés sont les attributs du co-domaine de la relation. Nous supposons que nous avons déclaré dans l'ontologie que toutes les relations sont deux-à-deux disjointes, excepté les deux relations *ContaminationLevelRelation* et *MaxContaminantionLevelRelation*. Nous supposons que tous les attributs sont deux-à-deux disjoints sauf les deux attributs *ContaminationLevel* et *MaxLevel*. Nous constituons, tout d'abord, l'ensemble de paires d'instances de relations comparables en tenant compte des contraintes de disjonction entre les relations de l'ontologie. Nous comparerons ainsi deux-à-deux, toutes les instances de la relation *ContaminationLevelRelation* du premier tableau, avec à la fois les instances de la même relation et celles de la relation *MaxContaminationLevelRelation* du second tableau. Pour illustrer le calcul de similarité entre deux instances de relations, nous détaillons ci-dessous le calcul de similarité de la paire d'instances : ($ContaminationLevelRelation_{13}$, $ContaminantionLevelRelation_{22}$). Leurs descriptions représentées sous la forme d'annotations floues sur leurs attributs sont les suivantes :

- {($Food_{13}$, {cereal bar/0.5, breakfast cereal sweet/0.40, cereal bar chocolat/0.40, cereal bar low calorie/0.35}), ($Contaminant_{13}$, {Ochratoxin A/1}), ($year_{13}$, [2003, 2003]), ($ContaminationLevel_{13}$,[0,0,0.2,0.2])}
- {($Food_{22}$, {breakfast cereal sweet/0.60, cereal bar/0.5, breakfast cake/0.5, cereal bar chocolat/0.40, cereal bar low calorie/0.35}),($Contaminant_{22}$, {Ochratoxin A/1}), ($ContaminationLevel_{22}$, [0,0,0.2,0.2])}

Nous calculons ensuite la similarité entre les instances des attributs comparables des deux relations. Pour calculer le score de similarité des instances d'attributs symboliques, nous utilisons la mesure de similarité $Jaccard_{flou}$: $score_{Inst}(Food_{13}, Food_{22}) = \frac{0.4+0.4+0.5+0.35}{0.6+0.5+0.4+0.5+0.35} = 0.7$, $score_{Inst}(Contaminant_{13}, Contaminant22) = 1$. Pour calculer le score de similarité des instances d'attributs numérique, nous utilisons la mesure de similarité présentée précédemment : $score_{Inst}(ContaminantionLevel_{13}, ContaminantionLevel_{22}) = 1$. Finalement, nous calculons le score final :

$score_{Final} = score_{Inst}(Food_{13}, Food_{22}) \times score_{Inst}(Contaminant_{13}, Contaminant_{22}) \times score_{Inst}(ContaminantionLevel_{13}, ContaminantionLevel_{22}) = 0.7 \times 1 \times 1 = 0.7$.

4 Conclusion

Le problème de redondance est primordial surtout dans le domaine de la gestion des entrepôts de données. En effet la redondance des données dans les sources peut les rendre inconsistantes, dégrade leur qualité et par conséquent leur exploitation. Le problème que nous avons abordé dans ce travail est la détection de redondances dans des tableaux extraits du Web et annotés sémantiquement à l'aide d'une ontologie de domaine. Notre approche s'inspire de la méthode de réconciliation N2R (Saïs et al. (2009)) dans la mesure où elle s'appuye sur des connaissances déclarées dans l'ontologie. Elle permet en plus de comparer les annotations

floues représentées dans les instances de relations. Des mesures de similarités entre ensembles flous ont été proposées pour réaliser ces comparaisons.

Références

Bilenko, M. et R. J. Mooney (2003). Adaptive duplicate detection using learnable string similarity measures. In *KDD*, pp. 39–48.

Chen, S.-J. et S.-M. Chen (2003). Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers. *IEEE 11*(1), 45–56.

Dong, X., A. Halevy, et J. Madhavan (2005). Reference reconciliation in complex information spaces. In *ACM SIGMOD 2005*, New York, NY, USA, pp. 85–96. ACM.

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition 5(2)*, 199–220.

Hignette, G., P. Buche, J. Dibie-Barthélemy, et O. Haemmerlé (2009). Fuzzy Annotation of Web Data Tables Driven by a Domain Ontology. In *ESWC*, pp. 638–653.

Jaccard, P. (1901). Bulletin de la société vaudoise des sciences naturelles. *37*, 241–272.

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

Tejada, S., C. A. Knoblock, et S. Minton (2001). Learning object identification rules for information integration. *Inf. Syst. 26*(8), 607–633.

Summary

We present a new method for detecting redundant data. It is applied to web data tables semantically annotated by an ontology. Our method uses ontology knowledge and computes similarity scores to decide the data redundancy. We have also proposed two similarity measures for numerical fuzzy sets as well as symbolic fuzzy sets.