Explication de décisions de réconciliation : approche fondée sur les réseaux de Petri colorés

Souhir Gahbiche*, Nathalie Pernelle*, Fatiha Saïs*

* LRI (CNRS & Université Paris-Sud) et INRIA Saclay
2-4 rue J. Monod, Parc-Club Orsay Université, F-91893 Orsay, France
{Souhir.Gahbiche@limsi.fr, Nathalie.Pernelle@lri.fr, Fatiha.Sais@lri.fr}

Résumé. L'objectif des systèmes d'intégration de données est de faciliter l'exploitation et l'interprétation d'informations hétérogènes provenant de différentes sources. Lorsque l'on doit intégrer de grands volumes de données, le recours à un expert n'est pas envisageable mais l'exploitation de processus d'intégration automatiques peut introduire des approximations ou des erreurs. Nous nous focalisons sur les résultats fournis par les méthodes de réconciliation de données. Ces dernières comparent les données entre elles et détectent celles qui réfèrent à la même entité du monde réel. Pour renforcer la confiance des utilisateurs dans les résultats retournés par ces méthodes, nous proposons dans cet article une approche d'explication graphique fondée sur les réseaux de Petri colorés qui est particulièrement adaptée aux approches de réconciliation globales, numériques et guidées par une ontologie.

1 Introduction

De nos jours, nous disposons de plus en plus d'informations provenant du web ou de différentes sources distantes. Ces informations sont créées à différents moments, par différentes personnes, pour répondre à divers besoins applicatifs, ce qui les rend inéluctablement hétérogènes. L'objectif des systèmes d'intégration de données est de faciliter l'exploitation et l'interprétation de ces informations hétérogènes en les intégrant dans un cadre uniforme de façon à donner l'illusion à l'utilisateur qu'il interroge une seule source. Pour réaliser de tels systèmes, il est possible de s'appuyer sur les ontologies afin d'intégrer les données sémantiquement. Les ontologies fournissent alors un vocabulaire structuré servant de support à la représentation des données et des requêtes.

Lorsque l'on s'intéresse à l'intégration sémantique de données, il faut faire face à deux problèmes de réconciliation. Le premier correspond au problème de réconciliation de schémas qui consiste à trouver des appariements entre les éléments (e.g concepts et relations) de deux schémas ou deux ontologies. Le deuxième problème est la réconciliation de données (ou réconciliation de références) qui consiste à mettre en correspondance des données provenant de différentes sources, et à détecter celles qui représentent la même entité du monde réel (e.g. la même personne, le même lieu, le même article). Ce problème est souvent critique. Ainsi, chaque année, de nombreuses entreprises doivent intégrer des milliers de références d'articles

provenant de catalogues fournisseurs vers le catalogue de l'entreprise. Les approches de réconciliation de données les plus informées sont les approches qui, d'une part exploitent les connaissances déclarées au niveau de l'ontologie et qui, d'autre part, sont dites globales : elles utilisent à la fois les attributs décrivant les données mais aussi d'autres données qui leur sont liées par les relations (Saïs et al. (2009); Dong et al. (2005)). Ainsi, une approche d'explication pour ce type de méthodes pourrait être aussi utilisée pour l'explication des résultats de méthodes de réconciliation moins complexes (i.e. locales, non itératives et n'exploitant pas les connaissances de l'ontologie).

Lorsque l'on doit intégrer sémantiquement de grands volumes de données, le recours à un expert humain n'est pas envisageable. Malheureusement, le remplacement de ces experts par des processus d'intégration automatiques introduit souvent des approximations ou des erreurs dans les données manipulées par les systèmes d'information et par conséquent dans les résultats des systèmes qui manipulent ces données. Dans le cadre d'un travail en collaboration avec des industriels sur la réconciliation des données, l'un des besoins exprimés par les utilisateurs des outils de réconciliation est l'explication de leurs décisions. Il est alors crucial de pouvoir renforcer la confiance des utilisateurs dans les résultats retournés. Ceci est possible en fournissant un moyen d'expliquer les résultats de réconciliation obtenus.

Nous présentons dans cet article un modèle d'explication adapté aux méthodes de réconciliation de données globales et guidées par une ontologie. Il permet de représenter le fait que les scores de similarité ainsi que les décisions de réconciliations soient interdépendants : la décision de réconciliation d'une paire de données influe sur la décision réconciliation d'une deuxième paire de données, qui elle-même peut influer sur une troisième et ainsi de suite. Nous prenons pour exemple la méthode N2R (Saïs et al. (2009)).

L'article est organisé comme suit. En Section 2, nous donnons quelques approches d'explication existantes, puis nous présentons la méthode N2R en section 3. Le modèle d'explication est détaillé en Section 4 et son implémentation en Section 5.

2 L'explication dans les systèmes d'intégration

La plupart des approches d'explication conservent et utilisent les traces des différentes opérations effectuées pour les exploiter lors de l'explication. Les approches (McGuinness et al. (2006)) et (Shvaiko et al. (2005)) supposent que les réconciliations peuvent être le résultat d'une approche purement logique. Ainsi, S-Match (Shvaiko et al. (2005)) traduit le problème de réconciliation en un problème de satisfaction de formules booléennes qui est ensuite résolu par un moteur d'inférence dédié au problème SAT. Des ressources externes comme Wordnet sont utilisées pour inférer un ensemble de synonymies. (Borgida et al. (2008)) a été réalisé pour expliquer le raisonnement dans DL-Lite. Dans de tels systèmes, l'explication exploite la preuve formelle de l'appariement. Malheureusement, une réconciliation de données ne peut pas toujours être découverte à l'aide d'un raisonnement purement logique lorsque les données sont hétérogènes.

Une approche telle que (Robin et al. (2004)) propose d'expliquer les résultats obtenus par la méthode iMAP qui calcule les scores de similarité de couples d'attributs dans l'objectif de réconcilier des schémas. L'explication facilite l'interaction avec l'utilisateur qui peut choisir le ou les meilleurs appariements parmi les candidats qui lui sont proposés, ce qui peut modifier le raisonnement effectué pour d'autres appariements. Dans cette approche l'explication fait

partie intégrante du processus de mise en correspondance. Différents éléments interviennent dans l'explication (contraintes, scores de similarité) mais cette méthode n'est pas adaptée à l'explication d'une réconciliation de données réalisée de manière globale et itérative. De plus, le volume de données à traiter est en général plus important pour les méthodes de réconciliation de données que pour les méthodes d'appariement de schémas. Il est alors important de différencier l'outil qui permet de détecter un ensemble de réconciliations de celui qui permet par la suite d'expliquer la similarité de deux données.

D'autre part, comme (McGuinness et al. (2006)), (Shvaiko et al. (2005)) et (Borgida et al. (2008)) l'ont fait dans un cadre logique, on peut s'intéresser à la présentation des explications en vue de l'adapter aux besoins et au type d'utilisateur.

3 Présentation de la méthode N2R

N2R est une méthode de réconciliation de références qui est globale et fondée sur un calcul de similarité entre les descriptions de données. Les décisions de réconciliation sont fondées sur les scores de similarité (valeur dans [0..1]) et un seuil de réconciliation donné (dans [0..1]).

Le modèle de données utilisé dans la méthode N2R est le langage OWL (Ontology Web Language) qui permet de déclarer un ensemble de classes (relations unaires) et un ensemble de propriétés (relations binaires). Dans ce modèle de données, on distingue deux types de propriétés : (i) les relations (en OWL abstractProperty) dont le domaine et le co-domaine sont des classes et (ii) les attributs (en OWL objectProperty) dont le domaine est une classe et le co-domaine est un type de données de base (e.g. Literal, Date).

Le score de similarité de chaque paire de références est calculé en fonction de leur description commune (i.e. valeurs des attributs et des relations communes). Pour déterminer l'importance des différents attributs et relations dans le calcul du score de similarité d'une paire de références, les propriétés Fonctionnelles et Inverses Fonctionnelles sont exploitées. Par exemple, la propriété *Located* qui lie des musées à des villes est fonctionnelle (et notée PF(Located)). Cela signifie qu'un musée est situé dans une seule ville. Lors du calcul de similarité, si deux musées sont très similaires, N2R infère que les deux villes où sont situés ces musées sont également très similaires. Les dépendances entre les similarités des paires de références et les paires de valeurs sont modélisés dans un système d'équations. Chaque équation exprime la similarité entre deux références en fonction des scores de similarité des paires de valeurs de leurs attributs communs, mais aussi en fonction des scores de similarité des paires de références qui leur sont liées par les relations communes.

Les scores de similarité des valeurs des attributs sont calculés par des mesures de similarité connues telles que celles présentées dans (Cohen et al. (2003)).

Pour chaque paire de références, le score de similarité est représenté par une *variable* x_i. Ce score est calculé par l'équation $x_i=f_i(X)$, avec $X=(x_1,x_2,...,x_n)$ où *n* est le nombre de paires de références. Les scores de similarité des valeurs de base $Sim_v(v1, v2)$ sont représentés par des *constantes*. Chaque équation est de la forme :

$$f_i(X)=\max(f_{i-df}(X),f_{i-ndf}(X))$$

La fonction $f_{i-df}(X)$ est une fonction qui combine les scores de similarité des paires de valeurs et des paires de références avec lesquelles la $i^{ème}$ paire de référence est en *dépendance fonctionnelle*. La fonction $f_{i-ndf}(X)$ permet de combiner les scores de similarité des paires

de valeurs, et des paires de références avec lesquelles la i^{ème} paire de référence *n'est pas en dépendance fonctionnelle*.

Dans le système d'équations, lorsque la dépendance n'est pas fonctionnelle, un poids, qui indique l'impact d'un score de similarité sur un autre, est utilisé (voir Saïs (2007)).

La résolution de ce système d'équations non linéaires est effectuée par une méthode itérative inspirée de la *méthode de Jacobi* (Golub et Loan (1996)), qui converge (Saïs (2007)).

Exemple Illustratif. Nous présentons sur un exemple la façon dont N2R modélise les dépendances entre similarités par un système d'équations en exploitant les connaissances du domaine. La figure 2 représente un ensemble de données provenant de deux sources S1 et S2. Ces données sont conformes au schéma OWL-DL représenté en figure 1 et décrivant le domaine des lieux culturels.

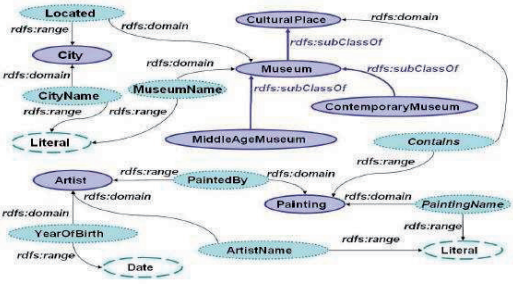


FIG. 1 – *Ontologie représentant les lieux culturels*

Source S1 :
MuseumName(S1_m1,"Le Louvre") ; Contains(S1_m1,S1_p1) ; Located(S1_m1,S1_c1) ; CityName(S1_c1,"Paris") ; PaintingName(S1_p1, "La Joconde") ;

Source S2 :
MuseumName(S2_m1,"musée du Louvre") ; Located(S2_m1,S2_c1) ; Contains(S2_m1,S2_p1) ; Contains(S2_m1,S2_p2) ;CityName(S2_c1,"Ville de paris") ; PaintingName(S2_p1, "Abricotiers en fleurs") ; PaintingName(S2_p2,"Joconde") ;

FIG. 2 – *Données RDF du domaine des lieux culturels*

L'ensemble des propriétés fonctionnelles est : {PF(Located), PF(YearOfBirth), PF(ArtistName), PF(PaintedBy), PF(PaintingName), PF(CityName), PF(MuseumName), PF(MuseumAddress)}, et l'ensemble des propriétés fonctionnelles inverses est {PFI(Contains), PFI(MuseumName)}.

L'exploitation des propriétés fonctionnelles et fonctionnelles inverses de l'ontologie et des scores de similarité des attributs permet de construire le système d'équations représenté dans Tab1. Chaque variable x_i représente le score de similarité entre références calculé par la fonction Sim_r et chaque constante représente le score de similarité entre valeurs calculé par Sim_v :

- $x_1 = Sim_r$(S1_m1, S2_m1) ; Sim_v("Le louvre", "Musee du louvre") = 0.68
- $x_2 = Sim_r$(S1_p1, S2_p1) ; Sim_v("La Joconde", "Abricotiers en fleurs") = 0.1

– x_3 = Sim$_r$(S1_p1, S2_p2) ; Sim$_v$("La Joconde", "Joconde") = 0.9
– x_4 = Sim$_r$(S1_c1, S2_c1) ; Sim$_v$("Paris", "Ville de Paris") = 0.42

Le tableau suivant contient le système d'équations obtenu avec les étapes de calcul :

Iterations	0	1	2	3	4
$x_1 = \max(0.68, x_2, x_3, \frac{1}{4} * x_4)$	0	0.68	0.9	0.9	0.9
$x_2 = \max(0.1, \frac{1}{2} * x_1)$	0	0.1	0.34	0.45	0.45
$x_3 = \max(0.9, \frac{1}{2} * x_1)$	0	0.9	0.9	0.9	0.9
$x_4 = \max(0.42, x_1)$	0	0.42	0.68	0.9	0.9

TAB. 1 – *Exemple d'un système d'équations et de sa résolution itérative.*

Le calcul du score de similarité pour chaque paire de références est itératif, et à chaque itération la valeur x_i est recalculée en utilisant les valeurs des variables (x_1, ..., x_n) de l'itération précédente. Les itérations s'arrêtent lorsque la différence des valeurs des variables entre deux itérations successives est inférieure ou égale à ε (point fixe à précision ε atteint). On considère alors que la solution du système d'équations représente les scores de similarité recherchés.

4 Représentation du calcul par un Réseau de Petri Coloré

L'objectif est d'expliquer à l'utilisateur un score de similarité obtenu par N2R en représentant graphiquement les différentes dépendances entre similarités ainsi que les différentes étapes de la résolution itérative du système d'équations. Nous proposons une approche d'explication fondée sur les Réseaux de Petri Colorés (RdPC). Il s'agit d'un formalisme qui permet de modéliser d'une façon lisible et compréhensible des systèmes à événements discrets et parallèles.

Dans cette section, nous allons tout d'abord présenter la transformation d'un système d'équations en un réseau de Petri coloré. Nous montrerons ensuite comment ce réseau est complété par des informations sémantiques afin de constituer une explication plus riche.

4.1 Transformation du système d'équations en un réseau de Petri Coloré

Un réseau de Petri classique est défini par un graphe où deux types de noeuds, les places et les transitions, sont liés par des arcs orientés. Les places peuvent contenir des jetons, représentant généralement des ressources disponibles. La distribution des jetons dans les places est appelée le marquage du réseau de Petri. Un réseau de Petri évolue lorsqu'une transition est franchie : des jetons sont retirés dans les places en entrée de cette transition et des jetons sont déposés dans les places en sortie de cette transition en obéissant aux règles de franchissement définies. Dans un Réseau de Petri Coloré (Jensen (1997)), on associe à chaque jeton une *couleur*. Cette couleur peut être de type complexe et permet de distinguer les jetons entre eux.

Le formalisme des réseaux de Petri colorés est utilisé comme un modèle d'explication graphique des décisions de réconciliations inférées par N2R. Transformer le système d'équations $F(X) = X$ en un réseau de Petri coloré consiste tout d'abord à créer pour chaque variable x_i et pour chaque constante c_i une place. Les scores de similarité des variables et des constantes

sont représentés dans la couleur des jetons de chaque place. Les jetons sont conservés permettant ainsi d'historiser les différentes valeurs des scores de similarité. Les couleurs des différents jetons sont définies comme suit :

Définition 1 ***(Couleur des jetons)***

Chaque jeton est caractérisé par une couleur de la forme (valeur, état).

- *La valeur représente le score de similarité, d'une paire de références (x_i) ou d'une paire de valeurs de base (c_j), représenté sous la forme d'un pourcentage.*
- *L'état a trois valeurs possibles :* old, new *ou* nc.
  - old *représente les jetons dont la valeur a été obtenue à une itération k-1 et qui doit être prise en compte pour calculer la nouvelle valeur de x_i à l'itération k. Pour les places représentant des constantes, la valeur de $etat$ est toujours à old.*
  - new *représente les jetons contenant les nouvelles valeurs de x_i à l'itération k,*
  - nc*, permet d'historiser les valeurs des jetons générés aux différentes itérations.*

Les dépendances entre les scores de similarité représentées par les équations sont modélisées dans le réseau de Petri résultat par des transitions et par des arcs liant les transitions aux différentes places. Le réseau est plus précisemment défini dans la définition (2).

Définition 2 ***(Le réseau de Petri coloré modélisant le système d'équations*** $F(X) = X$***)***

Le réseau de Petri coloré R_{rec} défini par le n-uplet (P, T, C, C_T, C_P, Pré, Post) représentant le système d'équations $F(X)$, où X=(x_1,...,x_n) utilisé dans N2R est défini par :

- *un ensemble fini de places $P=P_x \cup P_c \cup P_s$ où*
  - *P_x={p_{x1},..., p_{xn}} est l'ensemble de places attribuées aux variables de F*
  - *P_c={p_{c1},..., p_{cm}} est l'ensemble de places attribuées aux constantes de F*
- *un ensemble fini de transitions T= $T_e \cup T_s$ où T_e = {t_1, t_2, ...,t_n} avec t_i représentant l'équation f_i permettant de calculer la valeur x_i.*
- *un ensemble fini de couleurs C=($valeur_i$, $etat_j$)*
- *$\forall$ $t_i \in T_e$, $C_T(t_i)$=($valeur_i$, old) où $valeur_i \in$ [0..100], (couleurs des jetons acceptables par la transition).*
- *$\forall$ $p_{xi} \in P_x$, $C_P(p_{xi})$ = ($valeur_i$, $etat_j$) où $valeur_i \in$ [0..100], et $etat_j \in$ {old, new, nc}, (couleurs des jetons acceptables par les places représentant des variables).*
- *$\forall$ $p_{ci} \in P_c$, $C_P(p_{ci})$ = ($valeur_i$, old) où $valeur_i \in$ [0..100], (couleurs des jetons acceptables par les places représentant des constantes).*
- *à chaque variable x_j (respectivement constante c_j) apparaissant dans une équation associée à une variable x_i, un arc entrant allant de la place p_j vers la transition t_i est créé dans R_{rec}, et une précondition lui est associée : $\forall$ p_j, $\forall$ t_i, Pré(p_j, t_i) = ($valeur_i$, old) où $valeur_i \in$ [0..100], (couleurs des jetons acceptables par l'arc entrant). L'arc est bidirectionnel, i.e. les jetons sont consommés et remis dans la place.*
- *à chaque variable x_i pour laquelle une équation a été définie, un arc sortant allant de la transition t_i vers la place p_i est créé dans R, et une postcondition lui est associée : $\forall$ p_{xi} et t_i, Post(t_i, p_{xi}) = ($f_i(x_i)$, new) où $f_i(x_i)$ est la fonction de F qui calcule la valeur de x_i (couleurs des jetons calculés et associés à l'arc sortant)*
- *T_s, un ensemble de transitions additionnelles, associées aux places représentant des variables. Elles permettent de mémoriser des valeurs des variables calculées à une itération précédente Nous les avons nommées transitions de permutation.*

– P_s, *un ensemble de places permettant de contrôler l'ordre de franchissement des transitions de facon à franchir successivement les transitions représentant les équations puis celles qui représentent une permutation.*

Remarque : *Pour des raisons de clarté, dans la suite de l'article, les transitions T_s et les places P_s seront omises dans les réseaux de Petri colorés.*

Définition 3 ***(Marquage initial d'un réseau*** R_{rec}***)***

– *à chaque place représentant une variable, on ajoute un jeton de couleur (0,old).*
– *à chaque place représentant une constante, on ajoute un jeton de couleur (c_i, old) où c_i est le score de similarité obtenu par l'application des mesures de similarité de valeur de base utilisés dans N2R et définies dans (Cohen et al. (2003)).*

Exemple 1 ***Illustration de la transformation d'un système d'équations en un RdPC***

Considérons les deux équations suivantes appartenant au système d'équations présenté en section 3 :

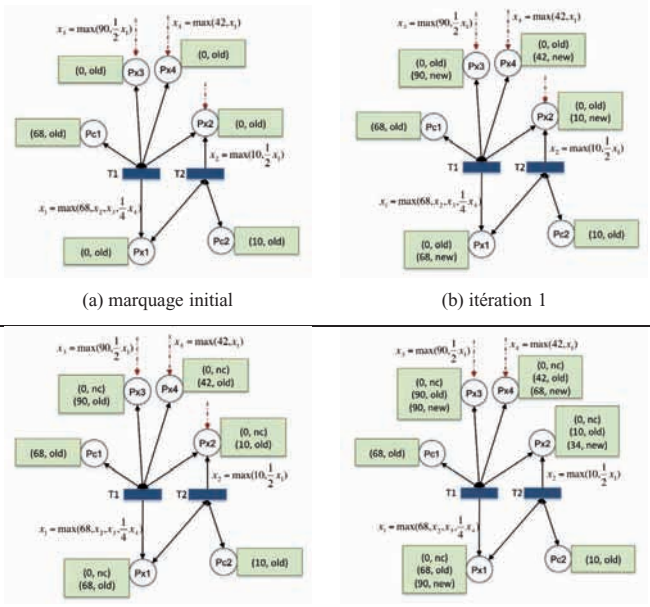
– $x_1 = max(c_1, x_2, x_3, x_4/4)$, *avec* $c_1 = 0,68$
– $x_2 = max(c_2, x_1/2)$, *avec* $c_2 = 0,1$

Pour obtenir le réseau de Petri Coloré correspondant R_{rec} donné en Figure 3.(a), nous créons :

– *six places P_{x1}, P_{x2}, P_{x3}, P_{x4}, P_{c1} et P_{c2} qui contiennent les jetons dont les valeurs sont initialement $x_1 = 0$, $x_2 = 0$, $x_3 = 0$, $x_4 = 0$, $c_1 = 10$ et $c_2 = 68$.*
– *une transition T1 qui prend en entrée les valeurs des jetons provenant de P_{x2}, P_{x3}, P_{x4} et P_{c1} et fournit en sortie le jeton allant à la place P_{x1}. Cette transition provoque le déclenchement du calcul de $f(x_1)$.*
– *une transition T2 qui prend en entrée les valeurs des jetons provenant de P_{x1} et P_{c2} et fournit en sortie le jeton allant à la place P_{x2}.*

Simulation du calcul itératif de similarité dans le réseau de Petri coloré. Le calcul itératif de similarité peut être simulé par les marquages successifs du réseau de Petri coloré R_{rec}.

Le réseau de Petri coloré de la Figure 3.(a) montre le marquage initial où l'on associe à chaque place représentant une variable, un jeton de couleur $(0, old)$ et à chaque place représentant une constante un jeton de couleur $(sim_v(v_1, v_2) * 100, old)$. Par exemple, le jeton déposé à la place P_{c1} est de couleur $(68, old)$ (68 correspondant au score de similarité de $(''Musee\ du\ Louvre'', ''Le\ Louvre'')$). Le franchissement de la transitions T_1 permet de calculer la valeur de x_1. La première itération du calcul est terminée lorsque toutes les transitions de T_e sont franchies. Les scores de similarité des valeurs de base ont alors été propagés aux places représentant les variables P_{xi} grâce à la création d'un nouveau jeton ayant comme état new et comme valeur, le score de similarité obtenu par l'évaluation de la fonction exprimée par $Post(t_i, p_{xi})$. Ensuite, une permutation est appliquée sur les couleurs des jetons des places P_{xi} : l'état des jeton old devient nc et celui des jetons new devient old. A la deuxième itération, les valeurs calculées à la première itération sont propagées, par la création de nouveaux jetons à état new dans chaque place. Le nombre d'itérations au bout duquel le point fixe est atteint est spécifié par un paramètre dont la valeur est donnée par N2R.



(a) marquage initial (b) itération 1

(c) permutation (d) itération 2

FIG. 3 – *Simulation du calcul similarité de N2R dans un Réseau de Petri Coloré (Extrait)*

4.2 Enrichissement du modèle d'explication

La vision structurelle et numérique du calcul de similarité est insuffisante pour qu'un utilisateur non spécialiste soit convaincu de la pertinence des scores de similarité et donc des décisions de réconciliation inférées. Ainsi, nous proposons d'enrichir ce modèle par des éléments d'explication supplémentaires : des éléments appartenant à la description des références (i.e. concepts, relations, attributs, et valeurs atomiques) mais aussi par des connaissances de l'ontologie qui permettent d'expliquer sémantiquement le calcul réalisé lors de la propagation des scores de similarité (i.e. propriétés (inverses) fonctionnelles).

4.2.1 Enrichissement par les éléments du schéma et par les valeurs atomiques

Nous proposons deux types d'enrichissement : un enrichissement par les éléments du schéma et un enrichissement par valeurs atomiques. Pour enrichir le modèle d'explication par **des éléments du schéma** OWL, nous proposons d'abord d'associer un label aux noeuds et aux arcs du réseau de Petri coloré. Cette première étape d'enrichissement est réalisée comme suit :

– A toute place du réseau représentant une variable (i.e. correspondant à une paire de références), ajouter un label exprimant le nom de la plus petite classe généralisant les classes auxquelles appartiennent les références comparées (e.g. Museum, Painting).

– A tout arc entrant du réseau liant une place représentant une variable et une transition représentant une équation, ajouter un label exprimant le nom de la relation commune [1].

– A tout arc entrant du réseau liant une place représentant une constante (i.e. correspondant à une paire de valeurs), et une transition représentant une équation, ajouter un label exprimant le nom de l'attribut commun.

Nous représentons également **les valeurs atomiques** associées aux références par les attributs communs (e.g. CityName(ref1, "Ville de Paris")). Il s'agit, en effet, d'une information essentielle pour l'utilisateur, car c'est à partir de ces valeurs atomiques que sont intialisés les scores de similarité des paires de références. Par exemple, en Figure 4, la place représentant la constante c_4 du système d'équations est étiquetée par le label *(Paris, Ville de Paris)* correspondant aux valeurs de l'attribut commun $CityName$ de la paire de références représentée par la variable x_4 dans le système d'équations.

4.2.2 Enrichissement par la sémantique

Dans N2R, le degré d'influence d'un score de similarité d'une paire de références sur une autre paire de références ou d'une paire de valeurs sur une paire de références dépend des connaissances déclarées au niveau de l'ontologie sur la fonctionnalité des relations et des attributs. Ces connaissances déterminent l'ensemble des dépendances fortes et faibles.

Pour distinguer ces types de dépendances dans le réseau de Petri, les arcs entrants des transitions représentant des équations sont colorés différemment. Les arcs à impact faible sont noirs et les arcs à impact fort sont colorés. Comme les propriétés fonctionnelles inverses peuvent être déclarées non seulement pour un attribut ou une relation mais pour des combinaisons de relations et d'attributs, nous créons alors des sous-groupes d'arcs à impact fort. A chaque sous-groupe, une nouvelle couleur (différente du noir) est associée.

Pour aider l'utilisateur à interpréter ces connaissances, nous associons aux PF et PFI une description textuelle traduisant leur sémantique. Par exemple, en figure 4 nous pourrions afficher à l'utilisateur à côté de la relation $Located$ qui est fonctionnelle le texte suivant "*Un musée est localisé dans une seule ville*".

4.3 Scénarii d'utilisation

Le modèle d'explication que nous avons développé peut être utilisé à différentes fins et pour différents types d'utilisateurs :

– Ces explications peuvent être utilisées comme outil d'aide à la décision lors de l'intégration de données dans un entrepôt ou lors de son nettoyage (élimination des données redondantes). L'explication permet alors d'aider l'utilisateur, formé à l'outil, à décider de la réconciliation ou de la non réconciliation d'une paire de références.

– Ces explications peuvent également être utilisées par les informaticiens en charge du système d'intégration pour détecter des anomalies dans les calculs de similarités. Des anomalies peuvent être dues à un mauvais choix de mesures de similarité pour les valeurs de base, ou à la présence de connaissances inexactes sur la fonctionnalité des propriétés dans l'ontologie. L'intérêt d'un tel modèle est d'être également un outil de simulation. Il est donc possible de l'utiliser pour tester les conséquences de certaines modifications du système.

1. les relations (resp. attributs) sont considéré(e)s comme étant commun(e)s s'il s'agit de la même relation (resp. même attribut) ou s'ils sont lié(e)s dans l'ontologie par un lien de subsomption (e.g. $SculptePar \preceq RealisePar$)

– Il permet aux systèmes d'intégration de justifier une réconciliation, permettant ainsi de renforcer la confiance de l'utilisateur dans une réponse constituée d'éléments provenant de différentes sources de données.

5 Implémentation dans l'environnement CPN-Tools

Pour tester notre approche, nous avons utilisé la plate-forme CPN-Tools (http ://www.daimi.au.dk/CPNtools) qui est un outil d'édition, de simulation et d'analyse de réseaux de Petri colorés. A partir d'un document XML traduisant le système d'équations de N2R, éventuellement enrichi par des connaissances de l'ontologie, un document au format CPN, représentant le réseau de Petri coloré correspondant, est automatiquement généré.

Nous avons appliqué notre outil d'explication sur deux types de données : (i) des données synthétiques du domaine des lieux culturels et (ii) des données réelles tirées du benchmark *Cora* traitant du domaine des publications scientifiques et sur lequel N2R a été validée.

5.1 Modèle d'explication sur les données synthétiques

La Figure 4 représente le modèle d'explication généré par notre outil pour le système d'équations montré dans Tab. 1.

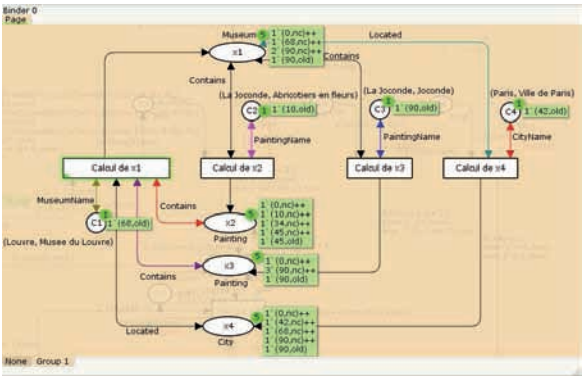


FIG. 4 – *Modèle d'explication pour N2R appliquée sur des données synthétiques*

Cette explication permet de montrer à l'utilsateur les dépendances entre les différents scores de similarité des paires de références et des paires de valeurs. Le franchissement des transitions montre la propagation des scores de similarité représentés dans les jetons. Dans un rectangle associé à chaque place, l'historique des scores calculés aux différentes itérations est présenté. Ainsi, la similarité x_1 est à 68 % puis à 90 %. Les connaissances sémantiques permettent d'informer l'utilisateur que, par exemple, la variable x_1 correspond à une paire de *Musées* et que leur similarité dépend d'une paire de villes et de deux paires de peintures. De plus, la paire de musées a un impact fort sur la similarité de la paire de villes (arc entrant de couleur verte allant de x_1 à la transition étiquetée "*Calcul de* x_4").

5.2 Modèle d'explication appliqué à des données réelles (Cora)

Nous avons appliqué notre approche sur les données Cora (6000 références d'article, de conférences et d'auteurs). Dans la Figure 5, nous montrons un exemple de paires d'articles décrits par leur titre, leur année de parution, leurs auteurs et la conférence où ils ont été présentés. Ces dernières sont décrites par leur nom, l'année, leur type et la ville où elles se sont déroulées. Nous avons déclaré au niveau de l'ontologie, les PFI suivantes : (i) pour la classe $Article$, $PFI(title, year, type)$ et (ii) pour la classe $Conference$, $PFI(confYear, confName, type)$. Cette représentation graphique de l'explication est adaptée à ce type de jeu de données : l'explication d'une paire de références implique en moyenne 5 variables et 5 constantes (donc un réseau de 10 places et de 5 transitions) ; N2R converge en moyenne en 10 itérations.

A-Ref	published	year	title	type	hasAuthors
441	440	1994	on-line prediction and conversion strategies	proceedings	436,437,438,439
412	411	1993	on-line predi ction and conversion strategies	proceedings	407,408,409,410

C-Ref	confYear	confName	type	city
440	1994	in computational learning theory : eurocolt '93 ...	proceedings	Oxford
411	1993	in computational learning theory : eurocolt '93 ...	proceedings	Oxford

FIG. 5 – *Exemple de données réelles (Cora) : articles, conférences et auteurs.*

L'application de N2R sur ces données a permis d'obtenir un score de similarité de 0.54 pour les articles (441, 412). Ce score est insufisant pour que N2R puisse prendre une décision de réconciliation ($<$ au seuil). Grâce à l'explication, l'utilisateur peut comprendre que ce faible score est dû à la dissimilarité des années. D'après les PFI déclarées dans l'ontologie, cette dissimilarité a un impact fort (à la baisse) sur le score de similarité des articles et des conférences. Néanmoins, en montrant les noms des conférences ("... *eurocolt'93*... ") l'utilisateur peut s'apercevoir que l'année "1994" associée à l'article 441 et la conférence 440 est erronée. Il peut donc suggérer la réconciliation de la paire d'articles mais aussi de la paire de conférences.

6 Conclusion

Dans cet article, nous avons proposé une approche d'explication des décisions de réconciliation obtenues par la méthode de réconciliation numérique N2R. Nous avons choisi et formalisé une approche fondée sur un outil de modélisation connu qu'est les réseaux de Petri colorés pour pouvoir générer un modèle d'explication graphique et extensible. Ainsi, nous pouvons montrer à l'utilisateur des explications lisibles comportant des connaissances sémantiques déclarées dans l'ontologie (e.g. les relations, les classes, les propriétés fonctionnelles).

Différents cas d'utilisation peuvent être définis pour cette approche d'explication : (i) renforcer la confiance de l'utilisateur dans les décisions de réconciliation de N2R en lui justifiant de leur pertinence ; (ii) consulter l'utilisateur pour les paires de données pour lesquelles N2R n'a pas pu fournir de résultat et (iii) permettre à l'utilisateur d'effectuer des diagnostics et de signaler des anomalies (e.g. erreurs dans les données, dans les connaissances, dans les choix de mesures de similarité).

Cette approche a été implémentée et testée sur des données synthétiques et réelles. L'outil a été développé en s'appuyant sur l'environnement de simulation CPN Tools.

Nous envisageons de tester notre approche d'explication sur d'autre jeux de données et d'effectuer une validation qualitative par les utilisateurs. Il serait également intéressant d'étudier l'applicabilité de notre approche pour l'explication de décisions obtenues par d'autres méthodes de réconciliation numériques de données ou de schémas.

Références

Borgida, A., D. Calvanese, et M. Rodriguez-Muro (2008). Explanation in the dl-lite family of description logics. In *Proc. of the 7th Int. Conf. on Ontologies, DataBases, and Applications of Semantics (ODBASE 2008)*, Volume 5332 of *Lecture Notes in Computer Science*, pp. 1440–1457. Springer.

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

Dong, X., A. Y. Halevy, et J. Madhavan (2005). Reference reconciliation in complex information spaces. In *SIGMOD Conference*, pp. 85–96.

Golub, G. H. et C. F. V. Loan (1996). *Matrix computations (3rd ed.)*. Baltimore, MD, USA : Johns Hopkins University Press.

Jensen, K. (1997). *Coloured Petri Nets, Basic Concepts*. Springer.

McGuinness, D. L., L. Ding, A. Glass, C. Chang, H. Zeng, et V. Furtado (2006). Explanation Interfaces for the Semantic Web : Issues and Models. In *3rd International Semantic Web User Interaction Workshop (SWUI06)*.

Robin, D., L. Yoonkyong, D. AnHai, H. Alon, et D. Pedro (2004). iMAP : discovering complex semantic matches between database schemas. In *SIGMOD '04 : Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp. 383–394. ACM.

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.

Saïs, F. (2007). *Intégration Sémantique de Données guidéée par une Ontologie*. Ph. D. thesis, Université Paris-Sud.

Shvaiko, P., F. Giunchiglia, P. P. da Silva, et D. L. McGuinness (2005). Web Explanations for Semantic Heterogeneity Discovery. In *ESWC*, pp. 303–317.

Summary

Data reconciliation consists in comparing data descriptions and detecting whether different descriptions refer to the same real world entity (e.g. person, place, gene). In order to enhance the user confidence in the results returned by the data reconciliation methods, we propose a graphical explanation approach based on coloured Petri nets. More specifically, the proposed approach is suitable with reconciliation approaches that are numerical, global and ontology driven.