EDA 2008, TOULOUSE

UNE MÉTHODE FLEXIBLE DE FUSION DE RÉFÉRENCES

Fatiha Saïs [1], Rallou Thomopoulos [1, 2]

[1] LIRMM (CNRS UMR 5506 et Université Montpellier II)
[2] INRA, UMR 1208

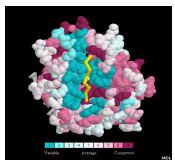
PLAN

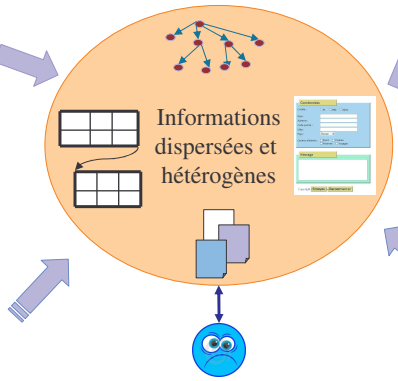
- Introduction
- Réconciliation de références
- Fusion de références :
  - Critères de classement des valeurs
  - Représentation floue des références fusionnées
  - Implémentation en RDF-Flou
  - Interrogation
- Conclusion et perspectives

INTÉGRATION D'INFORMATIONS, POURQUOI ?

Web

Sciences





Entreprises

Administrations

Intégrer et réconcilier ces informations.

DIFFÉRENTS PROBLÈMES DE RÉCONCILIATION

- **Réconciliation de données**
  - Détecter que deux descriptions de données représentent la même entité du monde réel.
- Réconciliation de schémas
  - Détecter que deux éléments de schémas représentent le même concept du monde réel.

RÉCONCILIATION DE RÉFÉRENCES

✧ Objectif : détecter que deux descriptions réfèrent à la même entité du monde réel.

Source 1

Ref	Nom	Rue	CP	Ville	Oeuvre
M1	Louvre	99, rue Rivoli	75001	Paris	La Joconde
M2	Arts premiers	37, quai Branly		Paris	
M3	…			…	…

Source 2

Ref	Nom	Rue	CP	Ville	Oeuvre
R92	Orsay	1, rue Légion d'Honneur	75007	Paris	
R50	Louvre	Palais royal	75001	Paris	Mona Lisa
R97	Quai Branly	37, quai Braly		Paris	

(M1 = ? R92) ;
(M1 = ? R50) ; …
(M2 = ? R92) ; …

Différents vocabulaires et conventions

Données erronées

Informations incomplètes

✧ Résultats : des décisions de réconciliation entre paires de références,
- ✧ Certaines : obtenues par des règles logiques [Saïs et al. AAAI'07]
- ✧ Potentielles : obtenues par un calcul de similarité numérique [Saïs et al. JoDS'08]

PLAN

- Introduction
- Réconciliation de références
- Fusion de références :
  - Critères de classement des valeurs
  - Représentation floue des références fusionnées
  - Implémentation en RDF-Flou
  - Interrogation
- Conclusion et perspectives

FUSION DE RÉFÉRENCES

✧ Objectif : fusionner les descriptions des références réconciliées pour pouvoir en obtenir une seule représentation.

① Réconciliation de (M1, R50)

M1	Louvre	99, rue Rivoli, 75001	Paris	La Joconde

Conflits

R50	Louvre	Palais Royal, 75001	Paris	Mona Lisa

② Réconciliation de (M1, Ref12)

M1	Louvre	99, rue Rivoli, 75001	Paris	La Joconde

Conflits

Ref12	Musée du Louvre	99, rue Rivoli, 75001		La Joconde

MÉTHODE DE FUSION DE RÉFÉRENCES

- Données du problème :
  - Soit R, un ensemble de n références $ref_1,\ldots, ref_n$.
  - Toute référence ref_i a une description :
    - $Desc(ref_i) = \{<ref_i\ A_1\ v_{i1}>, \ldots, <ref_i\ A_k\ v_{ik}>\}$
  - Des décisions de réconciliation potentielles entre références :
    - REC = $\{Reconcile(ref_1, ref'_1, s_{12}), \ldots, Reconcile(ref_i, ref'_j, s_{ij})\}$
- Résultat attendu :
  - Pour chaque attribut A_k, associer la liste de valeurs v_{ik} classées par un degré de confiance c_{ik} dans [0 ; 1].

CRITÈRES DE CLASSEMENT DES VALEURS

- Homogénéité des valeurs :

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

- Fréquence d'occurrence des valeurs :

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

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

- Fraîcheur de la source de données :

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

- Similarité globale de la paire de références : s_{ij}

DÉTERMINATION DU DEGRÉ DE CONFIANCE

- Soit A un attribut et $\{v_1,\ldots,v_p\}$ les valeurs respectives de *A* dans les descriptions des références $ref_1,\ldots, ref_n$ deux-à-deux réconciliées.

- Le degré de confiance *conf(v)* est obtenu comme suit :
  - Si *hom(v) = 1* alors *conf(v) =1* (v est la valeur de A pour toutes les références)
  - Si *hom(v) < 1* (v est la valeur de A pour certaines références)

$$conf(v) = max_{i\in I} \frac{Csim(v_i) + frch(S_i) + f(v_i)}{3}$$

REPRÉSENTATION EN RDF-FLOU

- Définition [Zadeh'65] :
  - *Un ensemble flou S, défini sur un ensemble de références X, est défini par une fonction d'appartenance μ_S de X dans [0 ; 1] qui associe à chaque élément x de X le degré $\mu_S(x)$ avec lequel x appartient à S.*
- Le langage RDF-Flou [Mazzieri'04] :
  - extension de RDF pour exprimer les triplets <sujet, prédicat, objet> par des déclarations d :<sujet, prédicat, objet>; avec
  - d est une valeur dans [0 ; 1] exprimant le degré de confiance que le couple (sujet, objet) soit une instance du prédicat.
- Descriptions de références fusionnées en RDF-Flou :
  - conf(v) : < ref_F, A, v>
  - Exemple : conf(''Louvre'') : <ref_{F1}, NomMusée, ''Louvre'' >

TRANSFORMATION DE RDF-FLOU EN RDF

- Transformation par réification :
  - Associer un identifiant *tripleID-i* pour chaque déclaration conf(v) : <ref_F, A, v>
  - Ajouter le prédicat confiance à la déclaration réifiée
  - La représentation réifiée en RDF est :

```
<tripleID-i    rdf:type          rdf:Statement > .
<tripleID-i    rdf:subject       myns:ref_F> .
<tripleID-i    rdf:predicate     myns:A> .
<tripleID-i    rdf:object        myns:v> .
<tripleID-i    myns:confiance    conf(v)^^xsd:decimal>  .
```

INTERROGATION

- Modes de fusion :
  1. Fusion totale : ne renvoyer qu'une seule valeur par attribut.
  2. Fusion partielle : renvoyer les k-premières valeurs par rapport au degré de confiance.
  3. Toutes les valeurs classées par degré de confiance.
- Exemple de requête SPARQL :

```
PREFIX  rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX myns: <http://www.lirmm.fr/~sais/myRDFS-1/>

SELECT  ?ref ?nom ?confiance
WHERE {
      ?x rdf:type rdf:Statement  .
      ?x rdf:subject ?ref  .
      ?x rdf:predicate myns:NomMusée .
      ?x rdf:object ?nom .
      ?x rdf:object ?confiance .}
ORDER BY ?confiance
LIMIT 1
```

CONCLUSION

- L'approche proposée :
  - Détection des conflits entre valeurs.
  - Représentation des descriptions de références fusionnées en utilisant des ensembles flous.
  - Différents modes de fusion à l'interrogation.
  - Implémentation en utilisant le langage RDF-Flou réifié sans l'extension du langage de requête.
- Approches existantes :
  - Ignorer les conflits [Papakonstantinou et al.'96]
  - Extension du langage de requête [Bleiholder et Naumann'05] et [Surahmanian et al.'95]

PERSPECTIVES

- Expérimentations sur des données réelles.
- Détection de conflits entre ensembles de valeurs.
- Une méthode globale de fusion de références.
- Interrogation des données fusionnées en exploitant les préférences de l'utilisateur.
- Gestion de profils utilisateurs en exploitant l'historique des requêtes.

MERCI POUR VOTRE ATTENTION !