

Stage de Master Recherche Informatique

« *Mais qui est mon Jean Dupont ?* »

Contribution au problème de l'identification d'individus à partir de descriptions sémantiques.

Encadrants : Michel Leclère (leclere@lirmm.fr) et Marie-Laure Mugnier (mugnier@lirmm.fr),
équipe RCR du LIRMM

Ce sujet est associé à une proposition de thèse pour septembre 2010. Il s'inscrit dans le cadre d'une collaboration avec l'ABES (Agence bibliographique de l'enseignement supérieur), basée à Montpellier, cf. site de l'ABES : <http://www.abes.fr>.

L'ABES gère les bases bibliographiques de l'enseignement supérieur (bibliothèques universitaires, thèses, ...). Chaque documentaliste, lorsqu'il saisit une nouvelle notice bibliographique (qui décrit le titre, les auteurs, le sujet etc... d'un document) doit vérifier si l'auteur, l'œuvre, et autres entités référencées dans la notice, existent déjà dans la base.

Pour cela, il fait une recherche par mot-clé dans la base. Pour les auteurs, par exemple, un simple nom ne suffit pas à les identifier de manière précise. Par ailleurs, un auteur peut avoir été identifié précédemment sous un nom différent pour diverses raisons (publication sous un « pseudo », mariage, faute de frappe, ...)

Le documentaliste doit donc sélectionner parmi les références retournées par la requête, celle qui correspond à l'entité qu'il veut référencer. Pour cela, il explore les descriptions des entités (par exemple pour les personnes : titre, nationalité, dates de naissance/mort,...) et les listes de documents précédemment attachés à cette entité.

En s'appuyant sur la description sémantique des entités d'une part et sur le réseau de relations entité/documents d'autre part, on peut chercher ce qui différencie chaque entité de toutes les autres entités de même nom.

Le but du stage est d'imaginer les grandes lignes d'un outil d'aide à l'identification des références existantes les plus pertinentes par rapport à la notice bibliographique en cours de saisie.

Plus précisément le stagiaire devra effectuer :

- Une étude bibliographique et comparaison des techniques proposées dans la littérature (syntaxiques ou sémantiques)
- Une première formalisation du problème posé s'appuyant sur les descriptions sémantiques des entités.
- Une proposition de scénarios d'application en collaboration avec l'ABES

Quelques références biblio sur le sujet :

- F. Saïs, N. Pernelle, M.-C. Rousset, Combining a Logical and a Numerical Method for Data Reconciliation, *Journal of Data Semantics*, vol. 12, p.66-94, 2009
<http://www.springerlink.com/content/y37943jr76x164t2/>
- O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S.E. Whang, J. Widom, Swoosh: a generic approach to entity resolution, *VLDB Journal*, vol. 18, n°1, p.255-276, 2009
<http://www.springerlink.com/content/72138t37t00jg554/>

- A. Elmagarmid, P. Ipeirotis, and V. Verykios, Duplicate Record Detection: A Survey, IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 19, n°1, January 2007
<http://pages.stern.nyu.edu/~panos/publications/tkde2007.pdf>
- A. Ferrara, D. Lorusso, S. Montanelli, Automatic Identity Recognition in The Semantic Web, IRSW 2008 <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-422/irsw2008-submission-2.pdf>

Quelques liens vers des projets en lien avec ce sujet :

- Le projet VIAF : projet d'unification des référentiels d'entités au niveau mondial
<http://www.viaf.org>
- Le projet SERF : projet de développement d'une infrastructure générique pour la résolution d'entité <http://infolab.stanford.edu/serf/>
- Le projet Sindice : projet de construction et exploitation d'un index des uri du web sémantique
<http://sindice.com/>
- Le projet Okkam : projet de construction d'un système de noms d'entité pour le web sémantique
<http://www.okkam.org/>

Sujet de thèse en liaison avec ce stage de recherche :

<http://www.lirmm.fr/RCR/theseidentification2009.pdf>