Méthode sémantique pour la classification et l'interrogation de sources de données biologiques

Nizar Messai, Marie-Dominique Devignes, Amedeo Napoli, Malika Smaïl-Tabbone

UMR 7503 LORIA, BP 239, F-54506 Vandoeuvre-Lès-Nancy, FRANCE
{messai,devignes,napoli,smail}@loria.fr
http://www.loria.fr/equipes/orpailleur

Résumé. Nous présentons une méthode de classification et de recherche de sources biologiques. Elle consiste à construire un treillis de Galois à partir d'un ensemble de méta-données associées aux sources et converties en propriétés booléennes. Un concept construit à partir d'une requête utilisateur est ensuite inséré dans le treillis grâce à un algorithme de construction incrémentale. Le calcul du résultat se ramène à extraire l'ensemble des sources figurant dans les extensions des subsumants du concept requête dans le treillis de Galois résultant. L'ordre de pertinence des sources est déduit à partir de l'ordre de subsumption des concepts correspondants dans le treillis. Une amélioration de la méthode consiste à enrichir la requête à partir d'ontologies de domaine avant de l'insérer dans le treillis. Deux modes d'enrichissement sont possibles: l'enrichissement par généralisation et l'enrichissement par spécialisation.

1 Introduction

Suite aux progrès accomplis dans la production et l'analyse de données biologiques, un grand nombre de données est rendu accessible via le Web. Ces données sont répertoriées dans des sources biologiques offrant des interfaces d'interrogation afin de faciliter l'accès à leurs contenus. La diversité de ces sources et la complémentarité des données qu'elles contiennent permettent aux utilisateurs d'avoir des informations plus complètes. Cependant, l'absence d'un schéma unique, l'incompatibilité des formats de données et l'absence (ou la faible fréquence) de mise à jour du contenu des sources peuvent entraîner des incohérences au niveau des réponses aux requêtes posées. Face à un tel problème, il peut se révéler utile de disposer d'une classification des sources selon des informations supplémentaires permettant de juger la pertinence des sources vis à vis des requêtes. Cette classification peut être faite sur la base d'un ensemble de critères documentant le contenu et la qualité des sources et appelés méta-données. À partir de la hiérarchie de sources obtenue, nous devons être capables d'extraire les sources susceptibles de répondre au mieux à une question donnée. La méthode d'interrogation des sources doit, en outre, prendre en compte la sémantique des requêtes qu'elle traite pour améliorer les résultats de la recherche. Ainsi le problème se ramène, d'une part à l'exploitation des connaissances (méta-données) décrivant les sources disponibles sur le Web dans le but d'identifier des sources pertinentes pour une question posée et d'autre part à l'analyse sémantique de la requête en se référant à des ontologies de domaine dans le but de raffiner cette requête et d'améliorer la réponse.