Classification et interrogation de sources de données biologiques

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

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.

2 Généralités sur les treillis de Galois

Nous ne rappelons ici qu'une définition générale des treillis de Galois. Des définitions plus détaillées peuvent être trouvées dans (Barbut 1970) et (Birkhoff 1973).

Un treillis est un ensemble ordonné tel que chaque couple d'éléments possède une borne *sup* (*supremum*) et une borne *inf* (*infimum*). Un treillis de Galois est le produit de deux treillis isomorphes appelés *treillis des extensions* et *treillis des intensions*. Les éléments du treillis sont des concepts. Un concept est formé d'un couple (*extension*, *intension*) où *extension* est un ensemble d'individus qui ont en commun l'ensemble des attributs *intension*. La relation d'ordre définie sur un treillis de Galois est appelée relation de subsumption.

3 Utilisation des treillis de Galois pour la classification et la recherche des sources de données biologiques

3.1 Construction incrémentale du treillis de Galois

En considérant la relation entre les sources et leurs propriétés comme une relation binaire entre un ensemble de départ (l'ensemble des sources) et un ensemble d'arrivée (l'ensemble des propriétés possédées par les sources), un treillis de Galois peut être construit. Nous nous retrouvons dans un cas similaire à la recherche documentaire en utilisant les treillis de Galois présentée dans (Godin et al. 1995) et (Carpineto 2000). Nous avons choisi "Incremental Structuring of Knowledge Bases" (Godin et al. 1995) comme algorithme de construction de treillis de Galois. Le choix de cet algorithme repose sur le fait qu'il offre la possibilité d'ajout de nouveaux concepts à un treillis déjà construit ce qui nous permettra d'une part l'insertion des requêtes dans le treillis pour récupérer les sources pertinentes et d'autre part l'ajout de nouvelles sources pour mettre à jour le treillis.

3.2 Recherche des sources pertinentes

Après avoir construit le treillis de Galois, un concept requête (dont l'intesion est formé de l'ensemble des propriétés de la requête) est inséré dans ce treillis. L'étape suivante consiste à chercher, dans le treillis résultant, le concept le plus spécifique contenant l'ensemble des propriétés exprimées dans la requête. On notera C_R ce concept. L'extension de C_R est l'ensemble des sources ayant toutes les propriétés exprimées dans la requête. À ce niveau, on ajoute au résultat à retourner (encore vide) les sources de cette classe qu'on notera R_0 (ensemble de réponses initiales de rang 0). On considère ensuite l'ensemble des concepts parents directs de C_R dans le treillis, appelés aussi subsumants les plus spécifiques de C_R Notons SS_1 cet ensemble. L'ensemble des réponses de rang 1, noté R_1, est formé des sources figurant dans les extensions des concepts de SS_1 sans considérer celles appartenant à R_0. De la même façon on détermine le reste des parties du résultat R_2, R_3, etc. jusqu'à atteindre le concept subsumant de C_R le plus général dans le treillis. Si l'intension de ce concept est vide alors aucune propriété n'est partagée par cet ensemble de sources avec la requête. Il n'y a donc plus de sources à ajouter au résultat d'où l'arrêt de la recherche.

4 Enrichissement sémantique de requêtes

4.1 Ontologies de domaine

La méthode de classification et de recherche de sources de données biologiques présentée ci-dessus est purement syntaxique et n'exploite pas les relations sémantiques susceptibles d'exister entre différentes propriétés. De fait dans la plupart des cas des relations sémantiques existent entre ces propriétés et sont exprimées dans des ontologies de domaine. Une ontologie est « *une spécification explicite et formelle d'une conceptualisation faisant l'objet d'un consensus* » (Gruber 1993). Elle réunit à la fois des éléments, concepts ou mots, et des règles permettant de manipuler ces éléments ou d'effectuer un certain nombre d'inférences (Berners-Lee 2001). La relation « *is-a* » définissant le lien de généralisation entre concepts est « *choisie comme relation de structuration de l'arborescence ontologique* » (Charlet et al. 2003). Nous représentons une ontologie par un arbre qu'on notera $T = (V, E)$ où V est l'ensemble des sommets de T qui représentent chacun une propriété appartenant ou non à l'ensemble des propriétés possédées par les sources et E est l'ensemble des arêtes entre les sommets de V. T est un arbre *enraciné* de racine r.

4.2 Enrichissement de la requête

Des méthodes de recherche d'information combinant les ontologies de domaine et les treillis de Galois ont été proposées dans (Priss 2000) et (Safar et al. 2004). Dans le premier cas, un thesaurus est utilisé pour enrichir l'indexation dans le treillis et améliorer le processus de recherche d'information. Dans le deuxième cas les ontologies de domaine sont utilisées pour construire un treillis raffiné selon les préférences de l'utilisateur en évitant ainsi la construction complète du treillis. Dans notre proposition, l'enrichissement consiste à ajouter à la requête utilisateur de nouvelles propriétés à partir des ontologies de domaine disponibles dans le but d'avoir un résultat plus riche. Il s'agit donc, pour une requête donnée, de considérer les propriétés figurant dans l'une des ontologies de domaine. Pour chacune de ces propriétés, on effectue un parcours de l'arbre T correspondant pour en extraire des sommets qui seront ajoutés en tant que propriétés à la requête initiale. On distingue deux types de parcours de T : un parcours par généralisation et un parcours par spécialisation. Ces deux modes correspondent respectivement à l'enrichissement par généralisation et à l'enrichissement par spécialisation.

4.2.1 Enrichissement par généralisation

L'enrichissement par généralisation consiste d'abord à localiser le sommet a de T correspondant à l'une des propriétés figurant dans la requête. Il faut ensuite parcourir le chemin de a jusqu'à la racine et ajouter à la requête les sommets rencontrés et qui appartiennent à l'ensemble des propriétés possédées par les sources. Cet enrichissement est effectué en particulier lorsque a est une feuille de T et que cette propriété n'apporte pas de source au résultat. Il permet d'obtenir une réponse enrichie par des sources plus générales que celles demandées par l'utilisateur vis-à-vis de la propriété appartenant à l'ontologie et figurant dans la requête initiale.

4.2.2 Enrichissement par spécialisation

L'enrichissement par spécialisation diffère de celui par généralisation par les sommets de T à ajouter à la requête initiale. En effet, au lieu d'ajouter les *ancêtres* du sommet a dans T qui figurent dans l'ensemble des propriétés possédées par les sources (a est le sommet qui correspond à l'une des propriétés de la requête), nous ajoutons les sommets du sous arbre $T' = (V', E')$ de T enraciné en a. Les descendants de a dans T représentent des spécialisations de la propriété représentée par le sommet a. L'ajout de ces propriétés à la requête se fait dans le but d'enrichir le résultat par des sources répondant à une partie de la requête considérée.

4.3 Apports de l'enrichissement sémantique de la requête

Après l'enrichissement de la requête, on effectue la recherche de sources pertinentes décrite dans la section 3.2 en considérant la nouvelle requête (enrichie). Le résultat de cette nouvelle recherche est plus riche en termes de nombre de sources pertinentes et l'ordre des sources dans le résultat est plus précis. En conclusion, l'enrichissement de la requête permet, en plus de l'enrichissement du résultat en nombre de sources, de réordonner plus précisément les sources selon leur pertinence et séparer éventuellement des sources qui avaient le même degré de pertinence (rang) dans le cas de la requête simple.

La pertinence des sources ajoutées découle du fait que celles-ci figurent dans le résultat grâce à une ou plusieurs propriétés ajoutées à la requête lors de l'enrichissement sémantique. Or, comme détaillé plus haut, tout ajout de propriété à la requête est dicté par une relation sémantique traduisant une certaine similarité entre les contenus des sources ayant la propriété ajoutée. De ce fait, le risque d'apparition de sources non pertinentes dans la réponse à retourner à l'utilisateur est écarté. Toutefois le degré de pertinence des sources ajoutées peut être faible.

5 Conclusion

Dans cet article nous avons étudié le problème de la classification et de l'identification des ressources biologiques pertinentes pour la réponse à une question donnée. La méthode que nous proposons s'appuie sur le fondement mathématique des treillis de Galois. Elle permet de construire une classification des sources sur la base du partage des propriétés et d'obtenir des réponses exactes lors de l'insertion d'une requête dans le treillis. Ce fondement mathématique permettra d'explorer de manière formelle le processus de recherche d'information tout en prouvant à chaque étape les résultats obtenus. La prise en compte de la sémantique valorise notre proposition vu l'intérêt grandissant accordé à cet aspect dans la majorité des domaines de recherche et notamment celui du Web sémantique.

Références

Barbut M.et Monjardet B. (1970), Ordre et classification : Algèbre et combinatoire (2 tomes), Hachette, Paris, 1970.

Berners-Lee T., Hendler J. et Lassila O. (2001), The Semantic Web, Scientific American, (284)5, pp 35-43, 2001.

Birkhoff G. (1973), Lattice Theory. American Mathematical Society Colloquium Publications, Rhode Island, 1973.

Carpineto C. et Romano G. (2000), Order-theoretical ranking, Journal of the American Society for Information Science, (51)7, pp 587-601, 2000.

Charlet J., Bachimont B. et Troncy R. (2003), Ontologies pour le Web sémantique, Action spécifique 32 CNRS/STIC Web sémantique Rapport final, Vol. 2, pp 43-63, 2003.

Godin R., Mineau G. et Missaoui R. (1995), Incremental Structuring of Knowledge Bases, Proceedings of the 1st International Symposium on Knowledge Retrieval, Use, and Storage for Efficiency (KRUSE'95), Santa Cruz (CA), USA, pp 179-193.

Godin R., Mineau G., Missaoui R. (1995), Méthodes de classification conceptuelle basées sur les treillis de Galois et applications, Revue d'intelligence artificielle, (9)2, pp 105-137, 1995.

Gruber T.R. (1993), A translation approach to portable ontology specification, Knowledge Acquisition, (5)2, pp 199-220, 1993.

Priss U. (2000), Lattice-based Information Retrieval, Knowledge Organization, (27)3, pp 132-142, 2000.

Safar B., Kefi H. et Reynaud C. (2004), OntoRefiner : a user query refinement interface usable for Semantic Web Portals, Proceedings of the ECAI 2004, Workshop on Application of Semantic Web Technologies to Web Communities, Valencia, Spain.

Summary

In this paper we present a method of biologic data sources classification and retrieval. It consists in building the Galois lattice of the binary relation between a set of data sources and their metadata converted to boolean properties. A Query concept can then be inserted into the Galois lattice using an incremental algorithm. The result is built with the set of data sources in the extensions of the query subsumers in the resulting Galois lattice. The ranking in the result is given by the subsumption order between the corresponding concepts in the lattice. The use of domain ontologies to refine the query before inserting it into the Galois lattice improves the retrieval performances. Two types of query refinement are possible : generalisation refinement and specialisation refinement.