Stage de Master 2ème année informatique ou 3ème année ingénieur

Représentation des connaissances et raisonnement pour la découverte de nouvelles connaissances

Contacts : claire.nedellecAT.inrae.fr, faronATi3s.unice.fr, liliana.ibanescuATagroparistech.fr

Mots clefs web sémantique, représentation des connaissances, raisonnement, ontologie

Date de début **mars 2021** ; durée de 6 mois ; stage rémunéré

Contexte

Dans les sciences expérimentales, les informations et connaissances du domaine sont décrites de différentes manières, e.g. en langue naturelle dans des publications scientifiques, et sous forme structurée dans des bases de données publiques ou privées.
L'origine des informations est une source de diversité. Les informations d'observations expérimentales sont précises et localisées, généralement numériques. L'extraction d'information à partir de textes (text mining) produit des informations de portée plus générale et de type qualitatif.
Leur annotation (ou indexation) par des ontologies facilite leur interrogation, leur accès et leur réutilisation [5,6]. Les modèles de représentations du web sémantique sont en effet adaptés à la gestion de très grands nombres d'observations ou de résultats expérimentaux décrits par des caractéristiques très variées.
Dans le domaine des sciences de la vie et de l'agriculture, la production de ces informations est coûteuse. Compléter ces données et produire des hypothèses réalistes par des moyens automatiques à partir de ces données est un enjeu majeur de ce domaine. L'indexation des données et des textes par des ontologies permet d'inférer de nouvelles informations. Le principe consiste à appliquer aux informations connues (observations, exemples) des règles du domaine, représentées dans un formalisme logique. Les inférences déductives sont les plus utilisées, puisque les assertions obtenues sont valides (par exemple, un canari est un oiseau). Les inférences non déductives, telles que l'induction, l'abduction ou l'analogie sont très intéressantes parce que les nouvelles assertions peuvent permettre d'enrichir considérablement les bases de connaissances, mais leur validité est conditionnée par la représentation des connaissances du domaine.
Ce stage s'inscrit dans le cadre du projet ANR D2KAB qui vise à créer une plateforme pour transformer des données en agronomie et biodiversité en connaissances - décrites de manière systématique, interopérables, exploitables, ouvertes - et étudier les méthodes et outils scientifiques permettant d'exploiter ces connaissances pour des applications en science et en agriculture.

Objectif

L'objectif du stage est de proposer une représentation formelle pour les connaissances du domaine et une méthode de raisonnement qui permette de déduire de nouvelles connaissances pour enrichir les bases de données.

Par exemple, on souhaite inférer dans la base Florilège, de nouvelles propriétés des microbes à partir de leurs habitats et des propriétés d'autres microbes.

Exemple : L'ontologie OntoBiotope définit la température interne du corps humain comme moyenne, et l'intestin comme faisant partie du corps humain. La littérature scientifique indique que la bactérie *E. Coli* vit dans l'intestin humain. On voudrait déduire de ces connaissances que la bactérie *E. Coli* peut vivre à température moyenne, elle est *mesophile*. On sait qu'une bactérie ne peut pas être à la fois thermophile (aimer le très chaud) et cryophile (aimer le très froid). On voudrait que les connaissances déduites respectent ces contraintes.

Comme dans cet exemple, on voudrait trouver et représenter des règles générales qui déduisent les phénotypes des organismes en fonction des propriétés connues des organismes et de leur environnment. On voudra aussi représenter des contraintes pour vérifier que les inférences sont cohérentes avec ces contraintes.

Les étapes du travail seront les suivantes :

- Étude bibliographique du raisonnement dans les ontologies et les données liées et comparaison des meilleures alternatives.
- Proposition d'une représentation formelle pour les connaissances de l'exemple Florilège et adaptable à d'autres sujets similaires (e.g. phénotypes du blé tendre)

- Proposition d'une représentation des contraintes et de leur vérification
- Réalisation d'une implémentation et évaluation expérimentale.

Les résultats feront l'objet d'une exploitation dans les bases de données des domaines expérimentaux considérés.

Méthodes, données et logiciels

Dans le cadre du stage, deux ensembles de données et ontologies d'INRAE seront considérés.

- La base publique Florilège (http://migale.jouy.inra.fr/Florilege/#&welcome) intègre des informations sur les microbes, leurs habitats et leurs phénotypes (leurs caractéristiques) provenant de la bibliographie et de bases de données biologiques. Ces informations sont indexées automatiquement par l'ontologie OntoBiotope (http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE/) et la taxinomie des espèces du NCBI (https://www.ncbi.nlm.nih.gov/taxonomy) [1].
- La base SamBlé intègre des informations sur les variétés de blé, leurs phénotypes et leurs traits provenant de la bibliographie et d'observations, indexées par l'ontologie WTO [5,7].

L'équipe Wimmics développe des modèles et outils basés sur les formalismes du web sémantique qui permettent de modéliser et de raisonner sur les ontologies et les données liées [4]. En particulier, le moteur Corese (https://project.inria.fr/corese/) permet (1) d'inférer de nouvelles connaissances à partir de sources de données RDF, en exploitant la sémantique de ces données capturée dans des vocabulaires RDFS, OWL ou SKOS ou des bases de règles d'inférence SPIN, (2) d'interroger ces données RDF en tenant compte de leur sémantique, (3) de vérifier la conformité des données par rapport à des contraintes exprimées en SHACL, et plus généralement (4) de traiter et visualiser des données RDF avec les langages LDScript [2] et STTL [3].

Lieu et encadrement

Ce stage sera réalisé dans le cadre d'une collaboration entre l'équipe Wimmics commune à Inria et I3S et deux équipes de deux unités INRAE et Université Paris-Saclay, l'équipe Bibliome de l'unité MaIAGE et l'équipe Ekinocs de l'unité MIA-Paris.

Lieu du stage : Unité MaIAGE, centre de recherche INRAE, Jouy-en-Josas
Encadrement :

- Claire Nédellec, équipe Bibliome, INRAE, Université Paris-Saclay, https://maiage.inrae.fr/fr/bibliome
- Catherine Faron, équipe Wimmics, Université Côte d'Azur, Inria, I3S https://team.inria.fr/wimmics/
- Liliana Ibanescu, équipe Ekinocs, MIA-Paris, INRAE AgroParisTech, Université Paris-Saclay https://www6.inrae.fr/mia-paris/Equipes/EkINocs

Références

1. Estelle Chaix, Louise Deléger, Robert Bossy, Claire Nédellec. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiology*, 2019. https://doi.org/10.1016/j.fm.2018.04.011
2. Olivier Corby, Catherine Faron Zucker, Fabien Gandon. LDScript: a Linked Data Script Language. *International Semantic Web Conference*, Oct 2017, Vienne, Austria.
3. Olivier Corby, Catherine Faron Zucker. STTL: A SPARQL-based Transformation Language for RDF. *11th International Conference on Web Information Systems and Technologies*, May 2015, Lisbon, Portugal.
4. Oumy Seye, Catherine Faron Zucker, Olivier Corby, Alban Gaignard. Publication, partage et réutilisation de règles sur le Web de données. *25èmes Journées francophones d'Ingénierie des Connaissances*, May 2014, Clermont-Ferrand, France.
5. Claire Nédellec, Liliana Ibanescu, Robert Bossy, Pierre Sourdille. WTO, an ontology for wheat traits and phenotypes in scientific publications. 18(2) *Genomics & Informatics*. 2020. doi: 10.5808/GI.2020.18.2.e1461.
6. Claire Nédellec, Robert Bossy, Estelle Chaix, Louise Deléger. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. In *Proceedings of the 4th International Microbial Diversity Conference*. pp. 221-227, ed. Marco Gobetti. Bari, October 2017. arXiv:1805.04107
7. Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, Pierre Sourdille. Information Extraction from Bibliography for Marker Assisted Selection in Wheat. In *proceedings of the 8th Metadata and Semantics Research Conference (MTSR'14),* Springer Communications in Computer and Information Science, Series Volume 478, Karlsruhe, pp 301-313, Allemagne, 2014. DOI: 10.1007/978-3-319-13674-5_28. https://hal.archives-ouvertes.fr/hal-01132767v1