AgroParisTech & INRA
UMR MIA 518
Equipe LINK
(Learning and INtegration of Knowldge)
16, rue Claude Bernard
75 231 Paris cedex 05

Université Paris Sud
LRI
Equipe LaHDAK

Sujet de stage de Master M2
Liage de données et transfert learning :
Comment apprendre un nouveau modèle d'un modèle existant en liant leurs données

Contacts : juliette.dibie@agroparistech.fr, cristina.manfredotti@agroparistech.fr, fatiha.sais@lri.fr

Objectifs :
L'objectif de ce projet est d'utiliser des méthodes de liage de données pour faciliter l'apprentissage d'un nouveau modèle de connaissances par transfert à partir d'un modèle existant « proche ». Il est à la croisée entre deux domaines de recherche en plein essor : le liage de données en Web sémantique et le transfer learning en apprentissage.
Le liage de données [REF1] consiste à détecter si des descriptions différentes réfèrent au même objet du monde réel. Les méthodes de liage de données s'appuient pour la plupart sur un calcul de similarité entre les données en utilisant des mesures de similarité élémentaires [REF2] connues dans la littérature, comme la mesure d'édition Levenstein pour la comparaison des valeurs atomiques. Lorsque les données sont décrites selon différentes ontologies, les méthodes de liage de données ont recours aux approches d'alignement d'ontologies dont l'objectif est de détecter les mises en correspondances entre les éléments (concepts, relations) des différentes ontologies. Par ailleurs, pour définir l'importance de telle ou telle partie de la description (propriété) des données, certaines approches [REF3] exploitent des connaissances, telles que les clés, déclarées dans les ontologies ou découvertes automatiquement [REF4] à partir des données. D'autres approches utilisent des techniques d'apprentissage supervisé ou non supervisé [REF5] pour apprendre les poids de certaines propriétés.
L'apprentissage par transfert (transfer learning) est une famille de méthodes d'apprentissage automatique qui s'intéressent à réutiliser des connaissances déjà acquises dans un domaine pour améliorer ou accélérer l'apprentissage dans de nouveaux domaines. Les travaux actuels sur le transfer learning proposent de passer par un espace de représentation commun dans lequel les fonctions de décision sont proches, voire similaires [REF6]. Cette approche présente des similarités avec le raisonnement par analogie [REF7] et peut être utilisée dans une perspective de choix de modèle descriptif commun à des données issues de différents domaines.
Nous proposons dans ce projet de développer une nouvelle approche de liage de données afin de traiter la question de modèles « proches » posée par le transfer learning en s'appuyant sur une comparaison entre les jeux de données sous-jacents aux modèles étudiés. L'idée est qu'en « liant » sémantiquement les données sur lesquelles s'appuient les modèles à rapprocher, nous pourrons utiliser ce liage pour faire le transfert. Dans notre cas, il ne s'agira pas d'un liage simple de données car les données à rapprocher sont analogues mais pas similaires. En guise de cas d'étude, nous considérerons dans ce projet le liage de données expérimentales décrivant le processus de stabilisation des levures et modélisées selon l'ontologie PO² [REF8] avec les données sur les recettes de cuisine (e.g., recette de cookies). En effet, le processus de stabilisation des levures comme la recette des cookies sont des processus de transformation composés d'une succession d'étapes qui permettent de faire passer un ou plusieurs objet(s) par différents états pour aller d'un état initial A à un état final B (e.g. les ingrédients de la recette des cookies comme la farine, les œufs ou le chocolat passent par différents états pour permettre finalement de fabriquer des cookies). On peut dire que leur modélisation est sémantiquement proche. Nous avons dans un précédent travail modélisé le processus de stabilisation des levures à l'aide du modèle relationnel probabiliste [REF9]. Le cas d'étude consistera donc à adapter ce modèle à la modélisation de la recette des cookies. Comme première piste de recherche dans ce projet nous envisageons d'étudier l'adaptation de mesures d'appariement de graphes pour définir une nouvelle approche de liage de données capable de détecter les descriptions de données analogues et sémantiquement proches. La comparaison des données proches devra se faire à un niveau conceptuel élevé en s'appuyant sur la structure des relations entre les données, le rapprochement sémantique concernant leur modélisation (e.g. les deux modélisations correspondant à des processus de transformation sont proches mais leurs domaines très différents). Nous proposons pour ce faire de combiner le liage de données et l'appariement de graphes sémantiques en s'appuyant sur l'ontologie du modèle source, tout en étant attentif à l'objectif final de transfer learning.

Résultats attendus, notamment en termes de développements informatiques:
L'objectif de ce projet est de développer une méthode permettant de comparer deux jeux de données dont les modélisations sont sémantiquement proches (e.g. des processus de transformations, des modèles de recommandation comme Netflix et Amazon) afin de faciliter le transfert du modèle connu et bien défini d'un des jeux de données à l'autre jeu de données sans avoir à tout réapprendre. A l'issue de ce projet les résultats escomptés sont :

- une étude bibliographique des méthodes d'appariement de graphes sémantiques
- une première méthode de liage de données capable de détecter les descriptions des données sémantiquement et

conceptuellement proches entre deux jeux de données de domaines différents. Cette méthode s'appuiera sur les techniques de comparaison de graphes sémantiques.
- Un prototype implémentant la nouvelle méthode de liage conceptuelle et sémantique des données.
- Une évaluation sur les données du cas d'étude, c'est-à-dire les données sur les processus de transformation de levures et les données sur les recettes de cuisine, ainsi qu'une première étude de faisabilité de prise en compte du résultat du liage de données afin de calculer la fonction de transfert du modèle relationnel probabiliste appris sur le premier jeu de données vers le second.

Les deux partenaires de ce projet sont d'une part l'équipe LInK de l'UMR MIA 518 experte en représentation des connaissances, construction et alignement d'ontologie et apprentissage, et, d'autre part, l'équipe LaHDAK du LRI experte en liage de données.
Les données manipulées lors du projet sont déjà disponibles : les données expérimentales décrivant le processus de stabilisation des levures et modélisées selon l'ontologie PO² ont été collectées et structurées par l'équipe LInK dans le cadre de divers projets (CellExtraDry, TransfOrm, LIONES, CARéDAS, NutriSensAl). Elles sont stockées dans la base de données relationnelle BaGaTel hébergée par la plateforme INRA PLASTIC. Des données sur les recettes de cuisine [REF10] sont disponibles en ligne (http://intoweb.loria.fr/taaable3ccc/).

Bibliographie

[REF1] A. Ferrara, A. Nikolov, F. Scharffe, Data Linking, In Web Semantics: Science, Services and Agents on the World Wide Web, Volume 23, 2013, Page 1, ISSN 1570-8268, https://doi.org/10.1016/j.websem.2013.11.001.
[REF2] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In Proceedings of the 2003 International Conference on Information Integration on the Web (IIWEB'03), Subbarao Kambhampati and Craig A. Knoblock (Eds.). AAAI Press 73-78.
[REF3] Saïs F., Pernelle N., Rousset MC. (2009) Combining a Logical and a Numerical Method for Data Reconciliation. In: Spaccapietra S. (eds) Journal on Data Semantics XII. Lecture Notes in Computer Science, vol 5480. Springer, Berlin, Heidelberg
[REF4] Symeonidou D., Armant V., Pernelle N., Saïs F. (2014) SAKey: Scalable Almost Key Discovery in RDF Data. In: Mika P. et al. (eds) The Semantic Web – ISWC 2014. ISWC 2014. Lecture Notes in Computer Science, vol 8796. Springer, Cham
[REF5] Nikolov A., d'Aquin M., Motta E. (2012) Unsupervised Learning of Link Discovery Configuration. In: Simperl E., Cimiano P., Polleres A., Corcho O., Presutti V. (eds) The Semantic Web: Research and Applications. ESWC 2012. Lecture Notes in Computer Science, vol 7295. Springer, Berlin, Heidelberg
[REF6] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. Proceedings of the 28th international conference on machine learning (ICML-11).
[REF7] Murena, Pierre-Alexandre, and Antoine Cornuéjols (2016). Minimum Description Length Principle applied to structure adaptation for classification under concept drift. 2016 International Joint Conference on Neural Networks (IJCNN). IEEE.
[REF8] L. Ibanescu, J. Dibie, S. Dervaux, E. Guichard, J. Raad (2016). PO2 - A Process and Observation Ontology in Food Science. Application to Dairy Gels. Proceedings of 10th International Conference on Metadata and Semantics Research Conference, MTSR 2016, pp. 155-165, Göttingen, Germany, November 2016.
[REF9] Melanie Munch, Pierre-Henri Wuillemin, Cristina Manfredotti, Juliette Dibie and Stéphane Dervaux (2017). Learning Probabilistic Relational Models using an Ontology of Transformation Processes. Proceedings of the 16th International Conference on Ontologies, DataBases, and Applications of Semantics, ODBASE 2017, Rhodes, Grece, octobre 2017, To appear ????
[REF10] Sylvie Després (2016). Construction d'une ontologie modulaire. Application au domaine de la cuisine numérique. Revue d'Intelligence Artificielle 30(5): 509-532

Lieu du stage: AgroParisTech (Paris), durée de 6 mois, stage rémunéré (environ 500 euros par mois)