Extraction d'expertise supportée par une ontologie pour la recommandation de flux de travaux scientifiques

Contexte - Une ontologie modélise un domaine ; elle en contient les concepts clés (ex. Ville, Musée), leurs rapports (ex. localisé-à), des instances de ceux-ci (ex. Washington, Smithsonian), etc. La fouille de données consiste à identifier toute régularité (tendance, concept ou motif récurrent) dans un grand jeu de données. La fouille de motifs [1] est une branche qui cherche les fragments intéressantes à partir d'une collection d'observations plus ou moins structurées (transaction, séquence, graphe, etc).

Problématique - Notre projet combine la fouille de motifs avec une représentation ontologique [4]. Les données sont des graphes acycliques orientés (DAG) étiquetés représentant des flux de travaux (workflows) élaborées à l'aide d'un outil bio-informatique pour l'analyse phylogénétique. Les noeuds sont des instances d'une ontologie du domaine phylogénétique (gènes, méthodes, outils logiciels, paramètres, etc.) Les arcs sont des liens entre noeuds de type entrée/sortie, paramètre-de, source-de, etc.

Nous visons la découverte de fragments de flux réutilisables –et donc abstraits– au sein d'une base de flux concrets construits par des usagers expérimentés. Pour en assurer l'abstraction des classes de l'ontologie sont employées en tant que noeuds dans les motifs. Ces motifs représentent l'expertise contenue dans la base de flux. Ils sont ensuite exploités par un outil de recommandation qui aide les utilisateurs novices à construire leurs propres flux.

Approche actuelle - Dans [2, 3] un cadre complet de fouille de motifs de workflows est décrit. L'approche générale de fouille consiste en ramener les DAGs vers une forme semblable à une séquence de transactions dont les items restent connectés via des propriétés (semblable à un tri topologique). Le but de la fouille est donc d'extraire les motifs de classe couvrant un nombre suffisant de DAG. Dans un second temps, un moteur de recommandation exploite ces motifs pour prédire la prochaine action à effectuer à partir d'un historique et de l'ensemble des motifs extraits.

Objectifs du stage - Solidifier les développements mathématiques qui sous-tendent la méthode de fouille. Développer des méthodes de recommandation utilisant les motifs généralisés de flux. Dans un deuxième temps, proposer de façons d'optimiser les mécanismes de la fouille.

Stratégie de réalisation - Le travail reprend les résultats d'une thèse de doctorat effectuée par A. Halioui [2]. Il s'agira de formaliser de façon complète les développements fondamentaux sur l'espace des motifs, les primitives et les algorithmes de fouille ainsi que de concevoir le moteur de recommandation (les techniques algorithmiques). Une validation formelle et expérimentale les mécanismes de fouille ainsi que de recommandation serait souhaitable. Les travaux seront accompagnés par un doctorant de l'UQAM, T.Martin et par A. Halioui.

Conditions de réalisation - La programmation est faite en Java et Python. La maîtrise de la notion d'ontologie serait souhaitable ainsi que des connaissances de base en fouille de motifs. La durée du stage est env. 5 à 6 mois.

References

[1] C. C. Aggarwal and J. Han. *Frequent Pattern Mining*. Springer, 2014 edition, Aug. 2014.

[2] A. Halioui. *Extraction de flux de travaux abstraits à partir des textes : application à la bioinformatique*. PhD thesis, Montréal (Québec, Canada), Université du Québec à Montréal, Doctorat en informatique cognitive, 2018.

[3] A. Halioui, T. Martin, P. Valtchev, and A. B. Diallo. Ontology-based workflow pattern mining: Application to bioinformatics expertise acquisition. In *Proceedings of the Symposium on Applied Computing*, pages 824–827. ACM, 2017.

[4] R. Missaoui, P. Valtchev, C. Djeraba, and M. Adda. Toward recommendation based on ontology-powered web-usage mining. *IEEE Internet Computing*, 11(4):45–52, 2007.