PROGRAMME INTITUTS ET INITIATIVES

Appel à projet – campagne 2021

Proposition de projet de recherche doctoral (PRD)

SCAI - Sorbonne Center of Artificial Intelligence

Intitulé du projet de recherche doctoral (PRD): Sémantisation de corpus scientifiques à large échelle - Application à l'analyse interactive de l'évolution des sciences.

Directeur.rice de thèse porteur.euse du projet (titulaire d'une HDR) :

NOM : **Amann** Prénom : **Bernd**
Titre : Professeur des Universités ou
e-mail : bernd.amann@lip6.fr
Adresse professionnelle : Sorbonne Université - LIP6 BC169 - 4, place Jussieu 75252 Paris cedec
(site, adresse, bât., bureau)

Unité de Recherche :
Intitulé : LIP6
Code *(ex. UMR xxxx)* : UMR 7606

École Doctorale de rattachement de l'équipe (future école doctorale du.de la doctorant.e) : **ED130-EDITE**

Doctorant.e.s actuellement encadré.e.s par la.e directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 1 - 2018 - 50%

Co-encadrant.e :

NOM : **Naacke** Prénom : **Hubert**
Titre : Maître de Conférences des Universités ou HDR ☐

e-mail : hubert.naacke@lip6.fr

Unité de Recherche :
Intitulé : LIP6
Code *(ex. UMR xxxx)* : UMR 7606

École Doctorale de rattachement : **ED130-EDITE**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de

doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) : 1 - 2018 - 50%

Co-encadrant.e :

NOM : Prénom :
Titre : Choisissez un élément : ou HDR ☐
e-mail :

Unité de Recherche :
Intitulé :
Code *(ex. UMR xxxx)* :

École Doctorale de rattachement : **Choisissez un élément :**
Ou si ED non Alliance SU :

Doctorant.e.s actuellement encadré.e.s par la.e co-directeur.rice de thèse (préciser le nombre de doctorant.e.s, leur année de 1e inscription et la quotité d'encadrement) :0

Cotutelle internationale : ☒ Non ☐ Oui, précisez Pays et Université :

Selon vous, ce projet est-il susceptible d'intéresser une autre Initiative ou un autre Institut ?
☒ Non ☐ Oui, précisez Choisissez l'institut ou l'initiative :

Description du projet de recherche doctoral *(en français ou en anglais) :*

Ce texte sera diffusé en ligne : il ne doit pas excéder 3 pages et est écrit en interligne simple.

Détailler le contexte, l'objectif scientifique, la justification de l'approche scientifique ainsi que l'adéquation à l'initiative/l'Institut.

Le cas échéant, préciser le rôle de chaque encadrant ainsi que les compétences scientifiques apportées. Indiquer les publications/productions des encadrants en lien avec le projet.
Préciser le profil d'étudiant(e) recherché.

* Problématique générale :

Aujourd'hui les graphes de connaissance (ou knowledge graph) sont en plein essor. Ils ont atteint un niveau de fiabilité très élevé et constituent une richesse immense pour mieux comprendre les données textuelles produites en masse chaque jour. Ils décrivent des connaissances en s'appuyant sur le standard RDF, et peuvent être interrogés à l'aide du langage de requêtes SPARQL. Un des plus grands graphes de connaissances est Wikidata qui apporte des informations encyclopédiques de culture générale sur de nombreux concepts et entités.

Toutefois, structurer des connaissances en RDF pour les intégrer dans un graphe de connaissance demande un effort important et on constate que de nombreuses connaissances sont actuellement produites sous la forme de documents textuels et ne se fondent pas sur le formalisme des graphes de connaissance. C'est le cas de la plupart des connaissances produites par les acteurs de la recherche scientifique au fil des années. Nous disposons ainsi de très grands corpus contenant des articles scientifiques publiés sur plusieurs décennies. Afin de faciliter l'exploration et l'analyse globale des connaissances représentées dans ces corpus, les approches existantes consistent à décrire (indexer) les documents par des ensembles de termes qui ne captent pas toutes les nuances conceptuelles nécessaires pour construire des cartes décrivant les interactions et évolutions scientifiques.

De ce fait, l'analyse plus sémantique des connaissances dans des corpus scientifiques se heurte à un manque de représentation sémantique des concepts scientifiques. L'objectif général de cette thèse est de proposer des nouveaux outils pour enrichir les grands corpus scientifiques avec des graphes de connaissances pour permettre une analyse plus fine des domaines scientifiques et de leur évolution. L'approche proposé consiste à combiner des méthodes de fouille de texte avec les technologies (RDF/SPARQL) et les ressources du web sémantique (Wikidata, DBPedia, Yago).

* Objectifs et défis scientifiques :

Les défis abordés seront :

- Défis n°1 : enrichissement sémantique des domaines issus de la fouille de texte : A partir des solutions de fouille de texte existantes qui calculent des domaines à partir d'un ensemble de documents, il s'agit d'enrichir les domaines en intégrant des données du web sémantique. En particulier, les domaines peuvent être représentés de manière plus riche qu'un ensemble de termes pondérés. Certains termes définissent le sujet du domaine, d'autres termes définissent la méthode

utilisée pour traiter le sujet, les acteurs impliqués dans l'étude scientifique et leur rôle, etc. Les annotations sont ensuite prises en compte pour calculer la similarité entre les domaines et caractériser la nature de l'évolution entre des domaines.

- Défi n°2 : recherche interactive de motifs d'évolution dans un grand graphe pondéré. Ce défi explore le traitement de requêtes complexes dans les graphes obtenus lors de la résolution du défi n°1. Analyser l'évolution des sciences consiste à rechercher un motif d'évolution dans un grand graphe où les nœuds sont des domaines scientifiques et les liens sont des canaux d'évolution. Le motif à rechercher est défini par une requête fixant la forme du sous-graphe à retrouver et la condition qui doit être satisfaite. Un motif d'évolution diffère d'une requête SPARQL classiques par deux aspects :

1) La condition nécessite généralement de calculer un agrégat qui dépend de l'ensemble des nœuds du sous-graphe à retrouver. Par exemple, retrouver les motifs où le degré l'évolution dépasse un seuil nécessite de calculer l'évolution moyenne des domaines inclus un sous graphe vis à vis d'un domaine de départ.

2) Un motif est souvent défini par par des opérateurs de type fermeture transitive : une instance du résultat doit contenir tous nœuds atteignables par les arcs dont la similarité est supérieure un seuil, sans toutefois que la valeur du seuil soit précisée dans la requête.

Une méthode simple pour évaluer une telle requête complexe consiste à partir d'un nœud racine quelconque, de parcourir progressivement les nœuds connexes jusqu'à obtenir un sous graphe qui satisfait la condition agrégative. Toutefois cette méthode ne passe pas à l'échelle : son coût devient prohibitif pour des grands graphes. D'autre part, nous considérons un cas d'usage interactif qui exige un faible temps de préparation des données à interroger. Il n'est donc pas envisageable pré-calculer à l'avance tous les motifs puis de les indexer.

Finalement, nous étudierons dans quelle mesure les algorithmes d'évaluation de requêtes proposés sont suffisamment généraux pour dépasser le cas de l'analyse de l'évolution des sciences.

* Justification de l'approche scientifique

Dans un premier temps une approche algébrique sera privilégiée. On étudiera la définition et l'implantation de nouvels opérateurs capturant la logique d'une jointure sémantique entre un graphe de connaissances et un ensemble de domaines décrits par des termes pondérés. L'accent sera également mis sur les requêtes qui effectuent des parcours transitifs dans les graphes. Des solutions seront étudiées pour optimiser l'analyse interactive, par exemple en anticipant les parcours demandés par les utilisateurs. Une représentation succincte du graphe sera étudiée afin d'améliorer les performances des requêtes.

Une approche expérimentale sera préconisée. Les algorithmes proposés seront implantés sur la plateforme Apache Spark en veillant à ce que leur exécution soit effectivement parallèle et distribuée et en améliorant, le cas échéant, le passage à l'échelle des traitements sous-jacents.

* Adéquation à l'Institut

Cette thèse s'inscrit dans le domaines des sciences de données et propose des nouvelles solutions d'analyse interactive de données complexes.

* Rôle et compétences scientifiques des encadrants

Les encadrants de cette thèse ont également un grande expérience de recherche en gestion et analyse de données à large échelle (Big Data, web), l'interrogation des données et l'optimisation de requêtes sémantiques. Ce travail de thèse s'inscrit dans la suite du projet interdisciplinaire ANR

EPIQUE (2018 - 2021) porté par l'équipe BD du LIP6 et intitulé « analyse de l'évolution des sciences à large échelle ». Le projet a produit des résultats sur le calcul efficace de la similarité entre les domaines [3], et la caractérisation des motifs d'évolution [1,2]. Le projet se termine prochainement et les nouveaux défis scientifiques énoncés plus haut restent ouverts.

Cette thèse bénéficiera également de l'expertise d'experts en philosophie des sciences de Institut d'Histoire et de Philosophie des Sciences et des Techniques (IHPST) et des collaborations avec l' Institut des Systèmes Complexes - Paris Ile-de-France (ISC-PIF).

* Profil de l'étudiant recherché

Nous recherchons un ou une candidate motivé/e avec des bonnes compétences en bases de données, algorithmique et programmation (Python, Java). Des connaissances en optimisation de requêtes et en traitement de documents et données sémantiques sont un plus.

* Bibliographie

[1] Ke Li, Hubert Naacke, Bernd Amann, EPIQUE: Extracting Meaningful Science Evolution Patterns from Large Document Archives. In International Conference on Extending Database Technology (EDBT, Demo). Copenhagen, Denmark, Mar 2020.

[2] Ke Li, Hubert Naacke, Bernd Amann, Exploring the Evolution of Science with Pivot Topic Graphs. In International Workshop on Big Data Visual Exploration and Analytics BigVis at EDBT 2020. Copenhague, Denmark, Mar 2020.

[3] Hubert Naacke, Ke Li, Bernd Amann, Olivier Curé, Efficient similarity-based alignment of temporally-situated graph nodes with Apache Spark. In IEEE International Conference on Big Data, High Performance Big Graph Data Management, Analysis, and Mining. Los Angeles, United States, Dec 2019.

[4] Hubert Naacke, Bernd Amann, Olivier Curé, SPARQL Graph Pattern Processing with Apache Spark. In GRADES (Graph Data-management Experiences & Systems), Workshop, SIGMOD 2017. pp. 1-7. Chicago, United States, May 2017.

Merci d'enregistrer votre fichier au format PDF et de le nommer :
«ACRONYME de l'initiative/institut – AAP 2021 – NOM Porteur.euse Projet »

Fichier envoyer simultanément par e-mail à l'ED de rattachement et au programme :
cd_instituts_et_initiatives@listes.upmc.fr avant le 20 février.