

PROPOSITION ACTION POUR LA FEDERATION LI-LIFO

1- Titre - *Semantic Big Data : querying, maintenance, evolution*

2- Description succincte (thématiques et objectifs)

Les propositions concernant les langages de requêtes sur des grandes masses de données sont nombreuses, mais des questions concernant le pouvoir d'expression, l'efficacité de l'évaluation, et l'assurance de la cohérence des réponses obtenues sont encore source de débat. Plusieurs questions concernant la manipulation, la maintenance ou l'interrogation d'un grand volume de données distribuées restent ouvertes : de la représentation et du choix de langages aux techniques de partage et équilibrage dynamiques des données ou de la charge de travail.

Dans le cadre des masses de données du web sémantique, appelées Linked Data (ou *Global Giant Graph*), les requêtes sur ces données prennent naturellement en compte leur sémantique. Dans ce contexte l'action s'intéresse à proposer des outils pour l'interrogation efficace (comportant des opérations d'agrégations ou autres) et déclarative de grands volumes de données distribuées organisées via un schéma contenant des informations sémantiques. Des contraintes d'intégrité et la confiance selon la provenance sont à prendre en compte. L'action s'intéresse aussi à l'évolution et l'intégration des contraintes de schéma et d'intégrité.

Toutes ces directions font le sujet d'une recherche en deux niveaux : front-end (langages de requêtes, algorithmes pour cohérence et intégration, etc) et back-end (algorithmes parallèles pour des requêtes complexes, distribution de données en fonction des types de requêtes plus fréquents).

3- Participants (personnes impliquées avec précision sur appartenance)

Joshua Amavi (LI/BdTIn) - ATER

Mostafa Bamha (LIFO/PAMDA)

Béatrice Bouchou-Markhoff (LI/BdTIn)

Jaques Chabin (LIFO/PAMDA)

Mirian Halfeld Ferrari Alves (LIFO/PAMDA)

4- Historique des collaborations (s'il existe des collaborations passées et des résultats déjà obtenus)

Des collaborations dans le passé dans le domaine de XML (mise à jour, validation de contraintes) et actuellement dans le cadre de la thèse Orléans-Tours.

5- Proposition de travail (description plus détaillée des collaborations envisagées et des résultats attendus)

Les collaborations ont déjà commencé à travers :

- Thèse Orléans-Tours : proposition d'un langage d'interrogation type datalog sur des données distribuées, algorithmes d'évaluation de requêtes prenant en compte la cohérence et la confiance de données.
- Participation au projet APR-IA, soumis récemment, et au projet européen H2020 I-Renaissance, en construction.

Des perspectives de collaborations internationales existent avec, en particulier, le Brésil et la Grèce.

Les premiers travaux se concentrent sur la question d'interrogation d'un graphe de données distribuées contraint par des règles d'intégrité qui doivent être vérifiées dans la production des résultats.

6- Prospectives (Interactions possibles avec autre action ? Auriez vous des idées sur les thématiques d'un axe qui pourraient englober cette action ?)

Des interactions sont déjà envisageables dans deux cadres :

- Création d'un langage de requêtes comme outil pour la préparation des données à analyser via des algorithmes de data mining, essentiels dans la découverte de connaissances sur de grands volumes données. Une interaction avec les actions concernant le data mining est donc nécessaire dans cette perspective.
- Efficacité des algorithmes d'évaluation des requêtes basés sur des environnements d'exécution distribués accompagnés des bibliothèques de programmation parallèle. L'interaction avec les collègues travaillant sur des algorithmes parallèles est donc souhaitable.

Un axe *Data Science* pourrait englober cette action.