Mission pour l'Interdisciplinarité du CNRS MASTODONS

Défi Grandes Masses de Données Scientifiques

CrEDIBLE

ConnaissancEs Distribuées en Imagerie BiomédicaLE

http://credible.i3s.unice.fr

Document ID. CrEDIBLE-12-3-v1

Décembre 2012

Résumé d'activité, année 2012

C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, J. Montagnat

Résumé

Ce document résume en une page le travail réalisé au cours de l'année 2012 dans le cadre du projet Crédible.

CrEDIBLE (ConnaissancEs Distribuées en Imagerie BiomédicaLE), Résumé 2012

1 Objectifs

La communauté médicale fait face à un éparpillement et un volume des données numériques croissant. Les cohortes de données accumulées constituent un capital précieux pour répondre aux défis actuels de la médecine puisqu'elles permettent de réaliser des études sur des populations à grande échelle, d'assurer un meilleur suivi des patients au cours du temps ou d'étudier des cas d'occurrence rare. Elles soulèvent cependant des défis considérables en raison de la quantité de données à manipuler, de la complexité des structures de données nécessaires à la représentation de l'information médicale et à la nécessité de disposer de référentiels et de métadonnées explicitant sa sémantique.

Le projet CrEDBILE s'inscrit dans une vision où les entrepôts de données médicales se trouveront de plus en plus systématiquement distribués, et où la capacité à fédérer ces entrepôts pour constituer, enrichir et traiter l'information prendra une importance croissante au cours des décennies à venir. La fédération de données dans ce sens recouvre (i) la fusion (virtuelle) d'entrepôts physiquement distribués, (ii) l'alignement sémantique de sources de données hétérogènes, (iii) la constitution de corpus de données d'étude par l'intermédiaire de requêtes, et (iv) l'analyse ou la transformation des ensembles de données ainsi constitués.

2 Travail réalisé

Le travail réalisé au cours des premiers mois du projet a consisté en une étude élargie des ressources ontologiques et des moteurs d'interrogation de bases de données sémantiques distribuées. Un prototype de moteur de recherche sémantique, basé sur le logiciel KGRAM a également été implanté.

Pour assurer une bonne prise en compte des travaux existants en relation avec le domaine médical et une bonne diffusion du travail réalisé, un atelier multidisciplinaire international a été organisé à Sophia Antipolis en octobre 2012. Il faisait intervenir des spécialistes des domaines de l'intégration des données, de la modélisation sémantique, des modèles de représentation de données, du raisonnement, et des flux de calcul sémantiques.

En s'appuyant à la fois sur l'expérience du projet NeuroLOG concernant la conceptualisation d'instruments de mesures utilisés en neurosciences et les présentations de travaux récents faites lors de l'atelier multidisciplinaire, un cadre ontologique a été défini pour rendre compte de la sémantique des données d'observation [rapport CrEDIBLE-12-3-v1, décembre 2012]. Ce cadre est organisé autour de DOLCE-CORE, une nouvelle version de l'ontologie fondatrice DOLCE développée au LOA (ISTC-CNR, Trento, Italie). Il étend DOLCE-CORE avec des modules couvrant les domaines des actions, des artefacts et des documents afin de mettre les données en relation avec d'autres entités participant de situations d'observations

L'interrogation conjointe de plusieurs sources de données, avec des capacités de jointure des résultats provenant de différents entrepôts, est un moyen de fédérer un ensemble de bases de données de manière transparente, en le faisant apparaître à l'interrogateur comme une seule base virtuelle. Les entrepôts de données biomédicales sont concernés par ces deux aspects puisqu'un nombre croissant d'études cherche à corroborer des informations de différentes natures (marqueurs biomédicaux issus de l'imagerie, tests cliniques, marqueurs biologiques…) ou à exploiter des bases de données contenant un information de même nature mais distribuées sur plusieurs centres d'acquisition (fédération de centres hospitaliers au niveau régional, création de cohortes de données de grande taille pour l'analyse statistique…). De nombreux travaux récents se sont concentrés sur la réécriture de requêtes et la génération de plans de requêtes permettant d'optimiser la performance du moteur d'interrogation distribué. Des compromis sont souvent réalisés concernant l'expressivité du langage d'interrogation accessible dans ce cadre. Les technologies du Web sémantique apportent cependant une grande flexibilité par la richesse du langage d'interrogation SPARQL d'une part, et la possibilité d'inférer des connaissances à travers l'application de règles déductives d'autre part. Une étude bibliographique détaillée des principales approches documentées dans la littérature a été réalisée. Un prototype de moteur de recherche dans des entrepôts sémantiques distribués compatible avec le standard SPARQL v1.1 est en cours de développement. Il permet l'interrogation de sources de données qui peuvent être représentées sous forme de graphes de connaissances (RDF, mais aussi graphes conceptuels, ou d'autres sources de données transformées telles que des bases relationnelles ou XML).

3 Bilan

Les travaux réalisés lors des premiers mois du projet CrEDIBLE ont conduit à une analyse approfondie du domaine de la fédération des données biomédicales réparties et des approches adoptées. L'atelier multidisciplinaire organisé répondait à la fois à des attentes réelles d'utilisateurs à la recherche de solutions pratiques et à une analyse des besoins de la communauté biomédicale d'un point de vu des concepteurs de méthodologies ou des développeurs de technologies. Fort de l'intérêt qu'il a suscité, une nouvelle session devrait être organisée courant 2013.

Le projet CrEDIBLE s'appuie sur une représentation sémantique des connaissances médicales (ontologies) et un moteur distribué d'interrogation et d'inférence sur des bases de connaissances qui implémente le standard SPARQL. Il permet de fédérer des bases de données distantes et hétérogènes via l'interrogation dynamique et la fusion de résultats provenant des différents entrepôts. Ce moteur s'adapte directement à l'interrogation de données RDF. Des travaux sont en cours pour développer une interface de médiation avec d'autres sources de données (notamment relationnelles).

[1] O. Corby *et al.* "KGRAM Versatile Inference and Query Engine for the Web of Linked Data", Web Intelligence, Macao, China, 12/2012.
[2] A. Gaignard *et al.* "Semantic Federation of Distributed Neurodata", DCICTIA-MICCAI 2012, pp 41-50, Nice, France, 10/2012.