Projet de recherche

Antoine Amarilli

De nombreuses sources de données structurées se développent aujourd'hui sur le Web : des bases de connaissances comme Wikidata, DBPedia, et YAGO ; des sources spécifiques comme OpenStreetMaps ou les données ouvertes de `data.gouv.fr` ; des annotations sémantiques sur des pages Web écrites dans les langages de Schema.org, Open Graph, etc. En utilisant ces données, on pourrait développer de nombreuses applications nouvelles : par exemple, apporter des réponses sémantiques à des questions structurées, comme le fait déjà Google avec ses Answer Box, et plus généralement offrir de nouveaux moyens de visualiser et d'intégrer ces sources de données.

Malheureusement, contrairement aux situations classiques en gestion de données, ces sources ne sont pas *fiables* : elles sont souvent créées ou extraites automatiquement par des règles faillibles, ou contribuées directement par les internautes. Les données peuvent ainsi être biaisées, incomplètes, voire périmées. On pourrait pourtant compléter les techniques actuelles de gestion pour ces données en utilisant des signaux liés à leur origine : les sites collaboratifs comme OpenStreetMaps ou Wikidata indiquent par exemple les différentes révisions de leurs données et les utilisateurs qui les ont créées, ainsi que leur *source*. Wikidata contient ainsi 35 millions de faits sourcés et OpenStreetMap 120 millions de voies sourcées.

Mon projet de recherche consiste à proposer une approche générale pour raisonner sur les données du Web en intégrant ces informations, en s'appuyant sur les techniques de gestion de la *provenance*. La provenance est une technique permettant de propager des annotations symboliques sur les résultats d'une requête, afin d'indiquer les sources et faits dont ils proviennent et la manière dont ils ont été calculés. La provenance a d'abord été définie pour les bases de données, où elle a fourni une solution générale à de nombreux problèmes auparavant étudiés indépendamment (gestion de politiques de sécurité, maintenance de vues, coûts d'accès, etc.) ; elle a été plus récemment étendue à d'autres domaines comme la gestion de données scientifiques ou les *flux de travaux*[1]. Mon projet de recherche consiste donc à *développer les fondements de la gestion d'une provenance expressive pour l'évaluation de requêtes et le raisonnement sur les sources de données du Web.*

1 Provenance et raisonnement

La principale difficulté pour définir et utiliser la provenance sur les données du Web provient du fait que la réponse aux requêtes sur le Web doit s'effectuer en *monde ouvert* : les données peuvent être incomplètes, et on souhaite évaluer des requêtes en déterminant leurs réponses certaines, en appliquant des techniques de *raisonnement* sous contraintes : logiques de description, règles existentielles, appariements de schémas[2], etc. La provenance, en revanche, a surtout été définie et étudiée pour l'instant dans le contexte de l'évaluation de requêtes relationnelles au sens classique ; c'est ainsi un défi majeur que de la généraliser au raisonnement logique sous des contraintes expressives. Mon projet de recherche consiste en premier lieu à développer une telle définition. Il vise également à étendre les approches actuelles pour le raisonnement, afin de calculer efficacement cette nouvelle notion de provenance ; et étudiera comment représenter cette provenance de façon concise, par exemple dans le formalisme récent des circuits de provenance, ou sous d'autres formes à développer.

Pour définir cette notion de provenance sur les données du Web, il faudrait étudier comment représenter les faits utilisés comme *hypothèses* pour le raisonnement, voire même les *règles logiques* utilisées pour aboutir

1. En anglais, *workflows*
2. En anglais, *schema mappings*

à une conclusion. On pourrait ainsi *comprendre* et *expliquer* les résultats du raisonnement en monde ouvert : l'utilisateur pourrait remonter aux sources et aux contraintes qui ont été utilisées pour déduire un résultat. Ceci nécessite une notion de provenance qui soit *abstraite* et indépendante de l'application visée, pour pouvoir la *spécialiser* ensuite, comme pour la provenance relationnelle à base de semianneaux. Ce projet est ambitieux mais toutefois réaliste : sa difficulté peut être modulée selon l'expressivité et le degré de généralité que l'on désire obtenir pour la provenance. Il répond également à un besoin pratique important. Par exemple, la base de connaissances YAGO, développée à Télécom par Fabian M. Suchanek en collaboration avec le MPI, maintient déjà une forme simple de *provenance* indiquant quelles règles d'extraction ont été utilisées pour chaque fait structuré. Mon projet proposerait une fondation logique pour généraliser de telles représentations.

2 Utilisation qualitative de la provenance

Une fois définie une notion de provenance pour annoter les résultats d'une requête en monde ouvert, il faut pouvoir l'utiliser pour en déduire des jugements sur les résultats. Un premier type de jugement est *qualitatif* : déterminer quel résultat est nécessairement meilleur que les autres résultats, si on sait que certaines sources sont meilleures. Notamment, si l'utilisateur a indiqué qu'une source était plus fiable qu'une autre, qu'on dispose d'un ordre temporel sur les données, ou d'un ordre de fiabilité sur les règles, il faut en déduire comment classer les résultats par ordre de pertinence pour l'utilisateur. On peut chercher à propager de tels jugements de pertinence de manière implicite sur les annotations de provenance, ou même envisager de raisonner sous des contraintes logiques qui peuvent utiliser cette relation d'ordre : "si deux faits sont incompatibles, préférer le plus ancien, sauf si l'utilisateur fait davantage confiance à une des sources qu'à l'autre".

Ce problème s'apparente au raisonnement en monde ouvert sous *relations d'ordre*, qui est un problème théorique délicat car la plupart des langages de contraintes décidables pour le raisonnement en monde ouvert ne permettent pas d'exprimer la transitivité de l'ordre. J'ai récemment commencé à étudier ces questions avec Michael Benedikt, Michael Vanden Boom (University of Oxford) et Pierre Bourhis (équipe LINKS INRIA Lille/CRIStAL), et envisage de poursuivre cette étude et d'en appliquer les résultats à la gestion de la provenance pour le raisonnement sur les données du Web.

3 Utilisation quantitative de la provenance

En plus des utilisations qualitatives visant à déterminer quels résultats sont plus pertinents que d'autres, la provenance sur un résultat peut également être utilisée de manière *quantitative*. L'utilisation la plus fréquente est d'adopter des modèles *probabilistes*, où les faits que l'on connaît sont annotés par des probabilités (estimées suivant leur source, ou par des techniques d'apprentissage ou de recherche de la vérité [3]). On cherche alors à calculer, pour chaque réponse à la requête, la probabilité qu'elle soit correcte.

Une telle utilisation probabiliste de la provenance pose cependant de nombreux défis de recherche. La définition d'une sémantique précise est déjà délicate, surtout si l'on veut aussi pouvoir indiquer que les *règles de raisonnement* sont également incertaines. Cependant, le principal problème est celui de l'efficacité : l'évaluation probabiliste est généralement hautement infaisable (#P-difficile). J'ai déjà étudié des méthodes pour rétablir la faisabilité [1, 2] dans ce contexte, et je compte m'intéresser à de nouvelles manières de faire cela dans le cas des données du Web, notamment en ayant recours à des techniques d'approximation et d'échantillonnage. Ce problème s'inscrit dans une collaboration avec Silviu Maniu (LRI) et Mikaël Monet, doctorant de Pierre Senellart à Télécom ParisTech qui travaille sur ces thèmes.

4 Retours utilisateur et révision

Lorsqu'on dispose d'annotations de provenance sur les résultats de notre raisonnement sur les données du Web, et que l'on peut les utiliser quantitativement et qualitativement pour estimer la qualité et la pertinence

3. En anglais, *truth finding*

des résultats à partir des données et contraintes initiales, la dernière étape consiste à intégrer les retours que fournissent les utilisateurs. Dans une vision applicative, on souhaiterait en effet que l'utilisateur puisse indiquer qu'un résultat est correct ou incorrect, soit directement, soit par des indices indirects comme le temps passé sur chaque résultat ou les clics effectués. On peut également vouloir utiliser des règles négatives comme des dépendances fonctionnelles pour savoir, par exemple, que deux résultats sont incompatibles et ne peuvent pas être vrais simultanément.

Ainsi, les perspectives à plus long terme de mon projet seraient d'utiliser de telles informations sur les *résultats* pour remonter, à travers la provenance, à des jugements sur les *données initiales*, notamment pour réviser la confiance qu'on leur accorde. Si on prend ainsi en compte la possibilité de solliciter des retours de la part de l'utilisateur, on peut même chercher à étudier quelles questions il sera le plus informatif de poser, ce qui se rapproche du problème d'interrogation de la foule [4] sur lequel j'ai déjà travaillé.

Intégration à l'équipe

Mon projet est en parfaite adhésion avec la thématique générale de l'équipe DBWeb, qui s'intéresse à la gestion de bases de données et aux données du Web. Sur le plan fondamental, en ce qui concerne la gestion de données incertaines, il correspond notamment à de nombreux thèmes précédemment étudiés par Pierre Senellart et son doctorant Mikaël Monet : Pierre Senellart m'a ainsi proposé de contribuer à l'encadrement de la thèse de Mikaël, qui a débuté en novembre 2015. Cependant, mon projet ne se résume pas à ces thématiques, et se concentre également sur d'autres notions : notamment, la gestion de la provenance, et le raisonnement sous contraintes logiques expressives. Ce sont là des sujets qui correspondent à mes propres intérêts, et à ma propre expertise.

Sur le plan applicatif, mon projet pourrait s'intégrer fructueusement au développement de l'ontologie YAGO par Fabian M. Suchanek à Télécom ParisTech. YAGO est une des sources de données structurées à laquelle on pourrait appliquer les méthodes de gestion de la provenance que je me propose de développer ; il serait particulièrement intéressant de chercher à appliquer ces notions de provenance pour des applications concrètes sur YAGO. On peut par exemple penser à l'intégration entre YAGO et d'autres sources de données comme Wikidata, i.e., le problème d'*alignement d'ontologies* sur lequel j'ai déjà travaillé. On peut également penser au raisonnement sur YAGO et à son extension avec des règles incertaines, que j'ai commencé à explorer avec Fabian et son doctorant Luis Galárraga. Enfin, on peut penser à l'intégration de sources de données externes avec YAGO, notamment les données du Big Data, et à l'extension et à la validation de YAGO avec la foule (une tâche déjà entreprise spécifiquement pour estimer le degré de fiabilité de YAGO).

Par ailleurs, les problématiques générales de gestion de la provenance et de l'incertitude, ainsi que le raisonnement avec les données, pourraient certainement trouver un cadre d'application dans les partenariats industriels dans lesquels l'équipe est active. Un exemple concret concerne les réseaux de transport, où la gestion de l'incertitude permet d'estimer une distribution statistique sur les temps de trajet : ce problème pourrait être étudié en collaboration avec SNCF et sa filiale Voyages-SNCF, dans le cadre de la chaire Big Data & Market Insights portée par Talel Abdessalem.

Collaborations extérieures

J'entretiens déjà différentes collaborations extérieures sur certaines directions de ce projet : avec Michael Benedikt, Michael Vanden Boom (University of Oxford) et Pierre Bourhis (CNRS CRIStAL) sur le raisonnement logique expressif ; avec Pierre Bourhis sur les méthodes structurelles pour l'évaluation probabiliste ; avec Yael Amsterdamer (Bar Ilan University, Israel) et Tova Milo (Tel Aviv University) sur l'interrogation de la foule ; avec Daniel Deutch (Tel Aviv University) et M. Lamine Ba (QCRI) sur les relations d'ordre incertaines.

4. En anglais, *crowdsourcing*