Collaborative Data-centric Workflows: Towards Knowledge centric workflows and Integrating Uncertain Data

Advisors : Pierre Bourhis (CR) - CNRS (CRIStAL), Serge Abiteboul (INRIA Paris), Lionel Seinturier (Université Lille)

Candidates interested in the subject must send a résumé and to Pierre Bourhis <pierre.bourhis@univ-lille1.fr>

The acquisition of data, in particular for scientific data, is more and more organized in complex processes that are captured by workflows. These workflows are often driven by ontologies. For example the collaborative application Spipoll [3] proposes to collect information about pollination in France. The users take pictures of insects on flowers, download them on the application and then identify the insect following a workflow based on a ontology characterizing the different insects. Finally, an expert certifies this characterization of the insect. Unfortunately, this certification takes time, as the number of experts is limited. Due to the lack of flexibility of the workflow, it is difficult to reuse data in other workflows or to adapt the workflow when needed. In particular, due to the lack in the management of uncertainty, managing efficiently the data acquired during the workflow is a difficult task. Indeed, data used in a workflow may be uncertain: because errors may be made during the experiments, or because scientists made mistakes. Such uncertain data may lead to wrong decisions. Moreover, because different people may have different interpretations, it may be the case that there is conflicting data. The goal of this PhD is to define a framework to deal with workflows that manipulate uncertain data and when data is acquired through knowledge-centric workflow.

The uncertainty of data may be capture in different ways:

> Errors in data acquisition. These can be expressed by different manners: probabilities, intervals of confidence. As data-centric workflows are defined by rules updating data, the uncertainty of data can be taken into account. This implies that the rules of the workflow can express properties over the confidence of data used to trigger the rules.
>
> Disagreements in data. The execution of a workflow may have to be revised. Moreover, when users disagree on some data, it is important to provide mechanisms to understand or select which data is correct for a particular user. Due to different choices, differences can be propagated to yield incompatible data living simultaneously in the same run.

The knowledge data, e.g. the ontology, can be used to guide the workflow to acquire some data or to determine some partial truth about data acquired even if there are some errors.

Approach

Model: First, we propose to formalize the different kinds of uncertainties previously mentioned in the context of collaborative data-centric workflows. For this, we will extend the notion of data centric workflow presented in [2] and we will try to capture workflows as described in Spipoll or in the demonstration in [1].

Confidence in data: We propose algorithms to deal with uncertainty: to compute efficiently a run with uncertain data and to deal with revision in data-centric workflows. We will study different tools to solve these problems. First, we will consider provenance that is information explaining where each piece of data comes from. Different works have shown that the provenance can be used to compute probabilities of an answer from a query over a probabilistic database; it can be used to show the impact of a revision of data on the result of a query over a database. We will study query answering in this setting. Second, we will introduce knowledge to capture contradictions, to improve the confidence in data, so that data can be shared more efficiently. Knowledge bases are very useful to integrate data. We plan to demonstrate that it turns to be useful to integrate data-centric workflows. For example, in the context of identifying insects on pictures, a data-centric workflow re-launches the identification if the picture is not surely identified for a set of families of insects. In a previous identification of the insects, the insect in a picture has been identified as a bee with a probability p and a bumblebee with a probability p'. However, both bees and bumblebees are part from the family of insects Hymenoptera. Therefore, the confidence that this insect is part of the family Hymenoptera is equal to 1. Then the data-centric workflow should not send this picture again for identification. Such problem will be formalized and studied during the PhD.

Application Finally, the techniques presented in this PhD will be validated in the context of the Spipoll activities.

The research activities associated with this PhD will benefit from the support and the collaborations that take place in the ongoing ANR funded project Headwork.

References

[1] Yael Amsterdamer, Susan B. Davidson, Tova Milo, Slava Novgorodov, Amit Somech: Ontology Assisted Crowd Mining. PVLDB 7(13): 1597-1600 (2014)

[2] Serge Abiteboul, Victor Vianu: Collaborative data-driven workflows: think global, act local. PODS 2013: 91-102

[3] http://spipoll.org