KEYSTONE COST ACTION IC1302, SHORT TERM SCIENTIFIC MISSIONS SCIENTIFIC REPORT

STSM Topic: Incremental Composite Entity Retrieval on the Web of Data
STSM Applicant: Khalid Belhajjame
Applicant's affiliation: Paris-Dauphine University, LAMSADE
Applicant's address: Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France

Purpose of the STSM

The objective of the STSM was to investigate retrieval of composite entities on the web that meet user needs by aggregating entities provided by multiple data sources. The problem of entity retrieval has been investigated by a number of researchers from the semantic web field. For example, FedX [SHH'11] is a system that allows users to setup federation of SPARQL endpoints that can be queried collectively by users. The focus in proposals, such as FedX, is on the optimisation of federated searches. In particular, users are assumed to be familiar with the relevant sources, and therefore, able to express federated queries using the terms (properties and eventually classes) of the sources.

The problem of heterogeneity in structure and semantics between the sources is to a certain extent ignored in federated search over the Web of Data. It was, however, partly addressed by proposals on (named) entity search [HMB'13]. Given a term that is provided by a user, the objective of named entity search is to identify the sources that contain the entity in question, which may have different representation in different sources. The assumption underlying the solution in this area is that there exist at least a source that contains the entity that the user wants.

In certain cases, however, the above assumption does not hold, in the sense that none of the sources contain the entity that the user is looking for. Instead, such an entity can only be obtained by retrieving and combining entities coming from different sources. In this STSM, we set out to investigate the problem of retrieving composite entities that meet user needs by aggregating entities provided by multiple data sources.

In the context of the visit of the applicant to the University of Manchester, we worked with the host (Pr. Norman W. Paton) and his collaborators from the University of Manchester (Dr. Alvaro A. A. fernandes and Mr. Duhai Ashukaili) to state clearly the problem and to identify elements of the solutions that can be developed further on after the STSM end.

Description of the work carried out and the main results of the the STSM

We started our work by trying to examine existing proposals in the semantic web and information retrieval on entity retrieval. In particular, we examined proposals on entity search, data integration and aggregated search from the information retrieval field. After several meetings, we came to the conclusion that there is a nish problem that has not been looked at, and formulated the problem of composite entity search on the Web.

My visit to Manchester and the interactions with the host and his collaborators was fruitful. Not only have we formulated and stated the problem more clearly, but we have looked at elements of the solutions that can be be further developed. In particular:

- We defined a query language for retrieving composite entities on the web.
- We investigated how user feedback can be leveraged for improving the quality of the search.
- We have identified the graph exploration techniques that can be used for effective exploration of the space of solutions.

Publication

We are currently working on the solution. The plan is that we have skype telecon every three weeks. Once the solution is matured and we empirically validated it, we plan to publish it in a semantic web or data management conference. We will of course acknowledge the support of the Keystone Cost action.

Future collaboration with the host institution

As stated above, we intend to go on collaborating with the host institution on the same topic. Besides the work on composite entity retrieval, I am now involved in a work of a PhD student at the University of Manchester (Duhai Al shukaili) who is working on a related topic, which is searching the semantic web using markov logic.

Confirmation by the host institution of the successful execution of the STSM

You will find enclosed a letter from the host (Norman W. Paton) confirming that the STSM was conducted successfully.