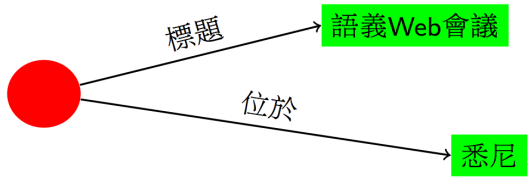
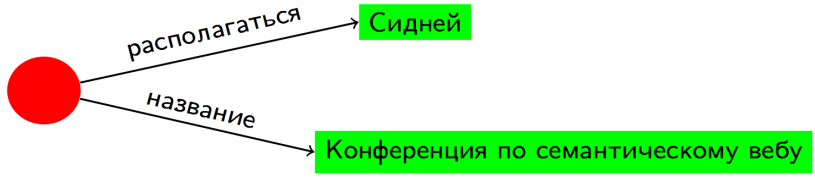
Multilingual Document Matching

The Semantic Web provides technologies such as RDF for representing data on the web. There is more and more initiatives making data available on the Semantic Web. For instance, DBpedia initiative represents Wikipedia content as RDF. In order to be used and useful such data has to be interlinked, i.e. the same entity (people, companies, events) represented in different data sets has to be identified. Data interlinking is a difficult task, particularly in a cross-lingual environment like the Web.

Consider the following two graphs:





Though they look similar structurally, it is not evident to say that they describe the same entity.

In that context, we are developing an approach which represents RDF entities as text documents and compare them using different strategies. One of them makes use of terminological resources like Wordnet[1]: the idea is to map important terms found in documents to the Wordnet and compute document similarity.

The goal of the project is to develop a software which can find documents that describe the same entity using specific methods.

Aim: Given the texts in two different languages, compute similarity between them and find the texts that describe the same entity (e.g. presidents, sportsmen, companies).

Data sources and technical details will be provided.

Technologies to be used:
Programming (Python, Java)
Knowledge of Natural Language Processing techniques and Computational linguistics is a plus.

Contact:

Tatiana Lesnikova
LIG & EXMO
tatiana.lesnikova@inria.fr

Jérôme David
INRIA
jerome.david@inria.fr

1 http://wordnet.princeton.edu/