Laboratory *LIRIS* - UMR 5205 CNRS ***Database group***

Internship for a Master 2 Research student - 2013-14

Semantic Enrichment of Entities in a Large Scale Context

Supervisors : Fabien Duchateau - Nicolas Lumineau
`firstname.lastname@liris.cnrs.fr`

Scientific Context

With the exponential growth of data, discovering relevant information about a given topic becomes a difficult task particularly when decision-makers have to quickly obtain relevant information to take the correct decisions. Thus, this internship aims at automatically building knowledges bases following an event (e.g., natural disaster, cultural event such as a new painting exhibition). The building of such knowlege bases is based on the enrichment of entities potentially relevant for the observation or the study of a geographical area, of a museum collection, etc. The multiplication of data sources from the Linked Open Data Cloud[1] (LOD) provides opportunities for integrating complementary or contradictory information. In addition, new types of data sources (e.g., blogs, Twitter) can also be exploited due to their frequent updates. A major problem for integrating these data sources on-the-fly lies in the large scale aspect.

Details of the Approach

The expected system aims at enabling users to publish entities of interest. These entities are enriched as more semantic correspondences are discovered with other data sources. In our context, two types of semantic enrichment occurs :

- 'Classic enrichment', which can be performed anytime (e.g., linking user's entities to their corresponding entities in the LOD cloud such as *DBpedia*[2] or *FreeBase*[3]). This process identifies the data sources which are interesting for a given user;
- 'Stream-based enrichment', in which an event is covered through streams of entities. In this case, the automatic extraction of semantic entities from the streams is performed automatically with time constraints (micro-blogging, blogs [1]). However, these extracted entities still have to be integrated with the entities of the 'classic enrichment'.

Our approach consists of defining a collaborative strategy for integrating all entities with acceptable performance.

[1] LOD, `http://www.linkeddata.org`
[2] DBPedia, `http://dbpedia.org`
[3] Freebase, `http://www.freebase.com`

Objectives and expected results

The objective of this internship is a definition of a collaborative strategy for processing the streams of semantic entities. The intuition is to rely on the "Map-Reduce" paradigm to pre-process the streaming entities. Expected results are a report about related work in distributed integration of semantic entities [2, 3], an original proposition for the collaborative strategy, and an experimental validation of the proposition.

Organization of the internship

This internship for a Master 2 Research student (5 months) is composed of three steps:

- Report on related work in distributed integration of semantic entities;
- Original proposition of a solution for processing the streams of semantic entities;
- Developement of a prototype to validate the proposition. Experiments should be run with real data on a large scale network simulator [4].

Complementary information

This research work is proposed in the *'database'* team of the LIRIS laboratory, as part of its activities in semantic data integration. The internship will be co-supervised by Fabien Duchateau and Nicolas Lumineau, both associate professors at LIRIS/UCBL. Fabien Duchateau notably works on semantic data integration [5]. Nicolas Lumineau notably works on collaborative alignment of ontologies [6].

References

[1] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *Proceedings of the 8th extended semantic web conference on The semanic web: research and applications - Volume Part II*, ESWC'11, pages 375–389, Berlin, Heidelberg, 2011. Springer-Verlag.

[2] Lars Kolb, Andreas Thor, and Erhard Rahm. Load balancing for mapreduce-based entity resolution. *CoRR*, abs/1108.1631, 2011.

[3] Aidan Hogan, Antoine Zimmermann, Juergen Umbrich, Axel Polleres, and Stefan Decker. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *Web Semantics: Science, Services and Agents on the World Wide Web*, 10(0), 2012.

[4] Alberto Montresor and Márk Jelasity. PeerSim: A scalable P2P simulator. In *Proc. of the 9th Int. Conference on Peer-to-Peer (P2P'09)*, pages 99–100, Seattle, WA, September 2009.

[5] Naimdjon Takhirov, Fabien Duchateau, and Trond Aalberg. An evidence-based verification approach to extract entities for knowledge base population. In *International Semantic Web Conference (ISWC)*, pages 575–590. Springer, 2012.

[6] Nicolas Lumineau and Lionel Médini. SimTOLE : un Simulateur P2P dédié à l'Alignement d'Ontologies à Large Echelle, January 2010. 10ième Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2010).