KEYSTONE COST ACTION IC1302, SHORT TERM SCIENTIFIC MISSIONS SCIENTIFIC REPORT

STSM Topic: Query answering in Evolving Datasets
STSM Applicant: Paolo Sottovia
Applicant's affiliation: Department of Information Engineering and Computer Science University of Trento
Applicant's address: Via Sommarive 9, 38123 Trento, Italy
Host's destination: Leibniz Universität Hannover Forschungszentrum L3S
Host's address: Appelstraße 9A, 30167 Hannover, Germany

Purpose of the STSM

The objective of the STSM was to investigate an entity centric query answering process in the context of the exploration of the web archives. This project is developed in the context of the Alexandria project (ERC Nr. 339233) that aims to develop models, tools and techniques necessary to explore and analyze Web archives in a meaningful way.
More precisely, we focused on the analysis and exploration of evolving entities. In this context one of the most important task to accomplish is the correct linkage of the entities during their lifespan.

The problem of entity linkage has been studied extensively for decades but only a few approaches consider data that evolves over time, where each data entry describes some aspects of a real-world entity at a specific time. These approaches assume that an entity can evolve only by changing its attributes over time. However, we believe that in many cases an entity not only changes its attributes but, over time, can also disappear or dissolve into various parts that may join other entities or may create completely new entities.

The mainly purpose of this STSM, in collaboration with prof. Wolfgang Nejdl and his PhD student Simon Gottschalk, is to formally define our problem statement related to an entity based query answering process over time. The other task is a plan for the future steps and goals in order to finalize this joint project.
Beside that, another important aspect about this STSM is the development of new research ideas involving researchers from both the institutions.

Description of the work carried out during the STSM

We started our work with an extensive analysis on the state of the art in the field of information retrieval, semantic web and data management. During the first week we had the opportunity to discuss in order to formalize the main problem and define the most important tasks that have to be done in order to complete the work.

During this discussion, one of main outcome was the need of a collection of temporal facts. The available knowledge bases such Freebase, Yago and Wikidata contain only few facts with temporal references.
In order to fill this gap, we defined a new approach in order to collect time-aware events from different sources. In particular we used the following sources:

- Wikipedia
- Wikidata

In the first phase we extracted all the Wikipedia infoboxes with event relative infobox types, then we use the Wikipedia category (e.g. establishments, birthdate and others) in order to extracted the lifespan of the entities. We also extracted all the event contained in some specific Wikipedia pages. Then, we linked these events with the related entities.
In the second phase we extracted all the outgoing links of a page and used these links in order to connect the entity related to a page with the event mentioned in the outgoing links. The pages related to an event has been identified with the infobox contained into the page or with the categories in which the page was inserted. We used the page hierarchy to extract more temporal information about the candidate events. In order to complete this work we need to integrate and resolve conflicts between the different events extracted.

In the remaining part of the SMST, we discussed about the next steps in order to continue the work after the end of the SMST. In particular, we decided how we can exploit the evolution of the entity, contained in the new knowledge base created, in order to apply it into a new query paradigm.

Description of the main results obtained

First, the major outcome of that STSM was the concrete establishment of a research collaboration between the two institutions.
Second, we had very good feedback from the event retrieval process. It is able to extract more that 500 thousand time aware events that are not present in the most known knowledge bases (Freebase, Yago and Wikidata).

Future collaboration with the host institution

As we stated before, we want to continue the collaboration with the host institution on the same topic, concluding the work and producing a scientific publication. In particular, we want continue the collaboration with the professor Wolfgang Nejdl and one of his PhD student Simon Gottschalk.

Publication

We are currently working on the project, we plan to continue it remotely and we plan to have a skype call every two weeks. Once the work will be complete we plan to publish it in a data management or semantic web conference.
We will of course acknowledge the support of the Keystone Cost action.

Trento, Italy, 22 April 2016

Paolo Sottovia