Towards Linked Data Extraction From Tweets

Manel Achichi*, Zohra Bellahsene*
Dino Ienco**, Konstantin Todorov*

*Université Montpellier 2, LIRMM
firstname.lastname@lirmm.fr,
**IRSTEA Montpellier, UMR TETIS
dino.ienco@teledetection.fr

Abstract. Millions of Twitter users post messages every day to communicate with other users in real time information about events that occur in their environment. Most of the studies on the content of tweets have focused on the detection of emerging topics. However, to the best of our knowledge, no approach has been proposed to create a knowledge base and enrich it automatically with information coming from tweets. The solution that we propose is composed of four main phases: topic identification, tweets classification, automatic summarization and creation of an RDF triplestore. The proposed approach is implemented in a system covering the entire sequence of processing steps from the collection of tweets written in English language (based on both trusted and crowd sources) to the creation of an RDF dataset anchored in DBpedia's namespace.

1 Introduction

One of the goals of the Linked Open Data (LOD) initiative is to structure and interconnect data on the web by using semantic web technologies, such as the Resource Description Framework (RDF), thus taking the web of today up to a new level where data are interpretable and accessible by both humans and machines (Bizer et al., 2009). A considerable effort has been made in that direction throughout the last couple of years. However, many sources of valuable information on the web still remain unexplored, although they contain useful data that can be beneficial for the LOD project. In this paper, we focus on the social medium Twitter that provides a platform for the publication of short messages (*tweets*) of maximal length of 140 characters. The network has enjoyed a growing popularity through the past years, becoming a major source of novel information about many important events, made available in real time, often even before its diffusion through the conventional broadcasting channels. The main motivation of our work is to enable the integration of the information flowing daily through the Twitter stream to the web of data. We propose a method for the extraction of relevant data from Tweets, their conversion into RDF and their storing into an RDF triplestore with the final objective of their publication as linked open data. Our approach covers the entire processing chain following a well-defined workflow, which will be presented in Section 3.