University of Fribourg / Faculty of Science and Medicine / Department of Informatics

Data Veracity Strategies to Evaluate Web Information Trustworthiness

Alex Carmine Olivieri

Data veracity is defined as the degree to which data is accurate, precise and trustable - the more the data fulfill these properties, the higher their data veracity rating. Data can have a low veracity because of different reasons, which span from precision errors to involuntary or voluntary human errors. The data veracity concept is becoming more and more important with the Internet evolution and the advent of Big Data and Open Linked Data, because these technologies allow data to be publicly shared between various actors. Moreover, the issue of data veracity is amplified by the emergence of social media and blogs, platforms on which anybody can publish information, thus increasing the possibility to have unreliable data. Providing methodologies that can verify data veracity is challenging but essential in today's world. Evaluating data veracity in the current open data world is a complex task because first of all it requires to contextualize the information, and then to verify if, for the domain at hand, the information provided by the data is correct. Current approaches focus on verifying the data for a specific domain, but fail to provide a methodology that can be applied and easily adapted for different domains. Furthermore, these approaches usually work offline, and need time to evaluate the veracity of the data - they are not responsive enough. The aim of this thesis is to propose a methodology for data veracity that can be easily adapted to various domains and that can be responsive enough to deal with the speed at which the Internet provides information nowadays. In this thesis, we describe several experiments with data veracity and the different strategies we used to study the issue. The experiments eventually led to the implementation of a methodology that uses search engines as a building block to find pieces of evidence, which are then evaluated using Machine Learning algorithms to perform the data veracity evaluation task. In this thesis we describe various approaches we applied to tackle this problem, which initially saw the involvement of humans, but turned to more automatic and sometimes hybrid approaches as the research progressed. We highlight the strengths of each approach, but we also acknowledge their weaknesses. We begin by illustrating some of the situations that highlighted the need for addressing the data veracity issue for a niche domain. We describe how we attempted to tackle this issue using people, passionate users first and then *expert* crowdsourcers. We dug more on crowdsourcing and how to select the most *expert* workers for a specific task given their answers to general test questions. As a result we produced guidelines for creating test questions for crowdsourcing workers selection. We then move our attention to automatic strategies for data veracity detection and we describe a system we developed that uses Semantic Web and Web Ontologies to create mappings of the domain for which we wanted to evaluate data veracity. We conclude by describing a framework that we developed which uses a Search Engine API in order to retrieve information about data that can be transformed into features that are then used by Machine Learning algorithms to classify their trustworthiness. This system can be easily adapted for different domains, and it can ideally run in real-time, which addresses the main gaps we identified in previous approaches for data veracity. Moreover, when testing this approach on a real dataset, the results of our experiments showed an improvement of accuracy of 3 % compared with the state-of-the-art approaches applied to the same dataset. The path to create efficient and effective methodologies for data veracity will still take time, mainly when dealing with unstructured data, but we strongly believe that the research described in this thesis can be used in combination with other studies, or as basis for others, in order to address data veracity issues in the future.

Jury:
Prof. Dr. Philippe Cudré-Mauroux, University of Fribourg (Switzerland), thesis supervisor.
Prof. Dr. Maria Sokhn, Neuchâtel University of Applied Sciences (Switzerland), thesis co-supervisor.
Prof. Dr. Heiko Schuldt, University of Basel (Switzerland), external examiner.
Prof. Dr. Jean Hennebert, Fribourg University of Applied Sciences (Switzerland), external examiner.
Prof. Dr. Edy Portmann, University of Fribourg (Switzerland), president of the jury.