Data Forensics with Analytics

Trust Discovery for Data Quality

Mouhamadou Lamine Ba and Laure Berti-Equille
Qatar Computing Research Institute

Data Forensics with Analytics, or DAFNA for short, is an ambitious project initiated within the Data Analytics Research Group in Qatar Computing Research Institute. It main goal is to provide effective algorithms and tools for determining the veracity of structured information when they originate from multiple untrustworthy sources. The ability to efficiently estimate the veracity of data, along with the reliability level of the sources in presence, is an ubiquitous problem in many real world use cases, e.g., data fusion or social data analysis, in which human rely on an automated data acquisition and integration process in order to consume high quality information for personal or business purposes. DAFNA's vision is to fill the gap of the lack of a comprehensive framework for information veracity management, with a direct impact on applications related to Qatar. Such a challenge requires to investigate various research topics such as a general purpose truth discovery algorithm and its applicability in practice.

We will present our ongoing study on extensively comparing the state-of-the-art truth discovery algorithms, releasing the first REST API for truth discovery algorithms, and designing an hybrid truth discovery approach using active ensembling. Finally, we will briefly present real applications of truth discovery in Qatar.

Efficient Truth Discovery Truth discovery is a hard problem to deal with in practical as often there is no a-priori about the veracity of provided information and the reliabily level of the sources. Thereby, the methodology followed by the majority of the existing truth discovery algorithms in order to determine the veracity of information is driven by the application domain and the characteristics of the dataset, e.g. the distribution of conflicts, the similarity of data item values, ect. This raises, therefore, some questions about a thorough understanding of the state-of-the-art truth discovery algorithms. A new truth discovery approach, if needed, should be rather comprehensible and domain independent. In addition, it should take advantage of the benefits of existing solutions, while being built on realistic assumptions for an easy use in real applications. We propose a study to deal with open truth discovery challenges which consists of the following initial contributions: (i) a thorough comparative study of twelve truth discovery algorithms; (ii) a design and release of the first REST API for truth discovery algorithms and; (iii) preliminary discussions about an hybrid truth discovery approach using active ensembling

Potential Applications in Qatar An obvious example of applications in Qatar which might take advantage of truth discovery algorithms is *Qatar Living* and similar systems. Truth discovery could help, indeed, to improve the quality of some information provided these applications. Given an increasing emergence of easy-to-use Web data extractors like *DeepDive*, one can imagine more general purpose Web applications for answering various types of requests about Qatar by simultaneously querying multiple sources such as Qatar Living, forums, blogs, social media, ect. We briefly propose a highlight of why truth discovery is incontournable in these contexts.