Discovering the truth in the Web Data

One Facet of Data Forensics

Mouhamadou Lamine Ba, Laure Berti-Equille, Hossam M. Hammady
Qatar Computing Research Institute, Hamad Bin Khalifa University

Data Forensics with Analytics, or DAFNA for short, is an ambitious project initiated by the Data Analytics Research Group in Qatar Computing Research Institute, Hamad Bin Khalifa University. Its main goal is to provide effective algorithms and tools for determining the veracity of structured information when they originate from multiple sources. The ability to efficiently estimate the veracity of data, along with the reliability level of the sources in presence, is a challenging problem in many real world use cases (e.g., data fusion, social data analytics, rumor detection, etc.) in which users rely on a semi-automatic data extraction and integration process in order to consume high quality information for personal or business purposes. DAFNA's vision is to provide a suite of tools for Data Forensics and investigates various research topics such as fact-checking and their practical applicability.

We will present our ongoing development (`dafna.qcri.org`) on extensively comparing the state-of-the-art truth discovery algorithms, releasing a new system and the first REST API for truth discovery algorithms, and designing an hybrid truth discovery approach using active ensembling. Finally, we will briefly present real-world applications of truth discovery from the Web data.

Efficient Truth Discovery Truth discovery is a hard problem to deal with in practical since there is no a priori knowledge about the veracity of provided information and the reliabily level of the sources. This raises many questions about a thorough understanding of the state-of-the-art truth discovery algorithms and their applicability for *actionable* truth discovery. A new truth discovery approach is needed and 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. In this context, we propose a study to deal with open truth discovery challenges and consists of the following contributions: (i) The thorough comparative study of existing truth discovery algorithms; (ii) The design and release of the first online truth discovery system and the first REST API for truth discovery available at `dafna.qcri.org`; (iii) An hybrid truth discovery approach using active ensembling; and (iv) An application to query answering related to Qatar where the veracity of information provided by multiple Web sources is estimated. []

References

[1] Berti-Equille Laure and Borge-Holthoefer J. *Veracity of Big Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics*. Morgane & Claypool Publishers. To appear.

[2] D.A. Waguih, N. Goel, H.M. Hammady, and L. Berti-Equille. Allegatortrack: Combining and reporting results of truth discovery from multi-source data. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 1440–1443, April 2015.

[3] Dalia Attia Waguih and Laure Berti-Equille. Truth discovery algorithms: An experimental evaluation. *CoRR*, abs/1409.6428, 2014.