Actionable Truth Discovery Challenges

LAURE BERTI-EQUILLE, MOUHAMADOU LAMINE BA, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.0 [**Database Management**]: General

General Terms: Algorithms, Management, Verification

Additional Key Words and Phrases: Truth Discovery, Fact-checking, Data Quality, Data Fusion, Information Extraction

ACM Reference Format:
Laure Berti-Equille, Mouhamadou Lamine Ba, 2015. Actionable Truth Discovery Challenges. *ACM J. Data Inform. Quality* 6, 6, Article 6 (November 2015), 4 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

As online user-generated content grows exponentially, the reliance on Web data and information from social networking is increasing in many domains for various private and corporate usages. One of the fundamental difficulties is that data can be biased, noisy, outdated, incorrect, misleading and thus unreliable. Massive data from multiple sources amplifies this problem since conflicting information have to be "aligned", compared and check to estimate their veracity. Obviously, truth discovery from the Web has significant practical importance: online rumor propagation [Kwon et al. 2013], mis- or disinformation can have tremendous impacts on our society, economy, politics, and homeland security. Online fact-checkers (e.g., FactCheck – http://www.factcheck.org/) and humanitarian initiatives [Imran et al. 2013] have rapidly appeared as unavoidable for classifying and verifying (semi-automatically) online information. Beyond social media and computational journalism, the truth discovery problem has been also largely studied in both the artificial intelligence and the database communities, sometimes under the names of "fact-checking" [Goasdoué et al. 2013], "information trustworthiness" [Thirunarayan et al. 2014], "information credibility" [Pasternack and Roth 2013], or "information corroboration" [Galland et al. 2010].

Given a set of assertions from multiple sources, the ultimate goal of truth discovery is to predict the truth label of each claimed data item and compute sources' reliability with no *a priori* knowledge about the truthfulness of the assertions and the sources. One major line of previous work, which relaxes this assumption, extended fact-finding models with prior knowledge either about the assertions [Pasternack and Roth 2013] or about the source reputation via trust assessment [Balakrishnan and Kambhampati 2011]. Another line of research aimed at iteratively computing and up-

Author's address: L. Berti-Equille, M. L. Ba, Qatar Computing Research Institute(Current address), Hamad Bin Khalifa University, Tornado Tower 18th, P.O. Box 5825, Doha, Qatar.

DOI: http://dx.doi.org/10.1145/0000000.0000000

dating the source's trustworthiness as a belief function in its claims, and then the belief score of each claim as a function of its sources' trustworthiness [Yin et al. 2008], with probabilistic models which incorporate domain-specific aspects like source dependence [Dong et al. 2010], evolving truth [Dong et al. 2009], hardness of certain claims [Galland et al. 2010], complex data structures [Zhao et al. 2012; Goasdoué et al. 2013], or multiple source's expertise [Ma et al. 2015]. However, current models suffer from various limitations, e.g., strong assumptions about claims and sources, restricting their usability to the wide diversity of Web information and realistic scenarios where knowing the truth as soon as possible is critical. We argue that the next-generation data-sharing systems need to manage not only heterogeneity but also unreliable information from various channels, in different languages, formats, and at different paces. Truth discovery systems have to be set up to help institutions and citizens by providing rigorous, scientific explanations and reports of their findings and collected evidences. Although some initial effort towards such a vision, e.g., [Dong and Srivastava 2013; Dong et al. 2010; Li et al. 2012], there is still a need for a more fundamental paradigm shift in data management to address the truth discovery problem natively. Formally, this goes beyond adding layers and extensions to data fusion heuristics, data provenance or truth discovery models. Technically, the challenges are not only to design techniques and prototypes of truth discovery systems but also to democratize operational tools for Web scale *information triage* and veracity verification.

2. ACTIONABLE TRUTH DISCOVERY CHALLENGES

We present a set of research actionable challenges that together aim at effectively address the truth discovery problem in real-world over-information scenarios.

Timely and Actionable Truth Discovery. Accessing past events and historical data in truth discovery is surely useful, however, from a humanitarian perspective, discovering truth from quasi real-time data could save lives. As a consequence, information extraction and triage as well as actionable truth discovery computation need to be streamlined, prioritized, and adjusted to the level of emergency and incompleteness of available information regarding the communities that will use it. This important usage-driven aspect goes beyond consistency checking and constraints.

Data Alignment Across Languages, Formats, and Channels. Agility of a truth discovery system is of utmost importance both at the technical, structural, and semantic levels. Agility is the ability of the system to efficiently extract and map information: (i) from various languages; (ii) various formats and data structures; and (iii) supported by various media and technologies. For instance, actual information extraction applications (e.g., Open-Calais – http://www.opencalais.com) are often specific to a single channel/application/language and cannot be considered as "agile". Moreover, each step of textual content analysis produces systematic and random errors that require to be considered in truth discovery unlike previous attempts like [Goasdoué et al. 2013]). Tracing and estimating the errors of data extraction, formatting, and linking is one of the main challenges for automating truth discovery.

Incomplete and Biased Observations. Observation data may be incomplete and biased for various reasons, e.g., security and privacy concerns, format limitations, and (unspecified) inferred values, leading to a truth discovery context where Open World Assumption must hold. Modeling the data provider's/observer's is crucial since online sources have various levels of *a priori* knowledge about the observation data collection technique, opportunistic data are *biased by observation effort*, and the underlying observation method is unknown. In addition, information may also suffer from *observer's*

bias and *disclosure bias*. One challenge is thus to estimate the incompleteness and biases and take them into account in the truth discovery computation.

Decontextualization and Distortion. Information without context can be easily distorted and misinterpreted. Indeed, when a piece of information is extracted from its original content and channel, it may lose its context along with important "semantic markers" that explain *when, where, how, why,* and *for which purpose/audience* it has been produced. Moreover, the pace has to be taken into account when information updates from various sources have to be *aligned* with considering the source granularity (e.g., an individual or an organization). Despites much research effort [Bao et al. 2010] in modeling contexts, current information extraction techniques constraint the context representation and can usually not handle a wide, unpredictable range of topics. Ultimately, all meta-information about the context and sources' characteristics have to be encoded as evidences and analyzed for truth discovery computation.

An interesting research direction to tackle these appealing truth discovery challenges could be an integrative framework. We believe that such a framework has to (i) to define, in a principled way, a concrete and unified semantics of truth discovery systems, e.g., using modal logics via axioms and canonical *Kripke* structures [Goranko and Otto 2006]; (ii) to support continuous inferrence and belief revision from new evidences, contextual data, and knowledge, inspired by recent development in continuous cleaning [Volkovs et al. 2014]; (iii) and finally to monitor and report uncertainty generated by the truth discovery process itself during information processing.

REFERENCES

Raju Balakrishnan and Subbarao Kambhampati. 2011. SourceRank: Relevance and Trust Assessment for Deep Web Sources based on Inter-source Agreement. In *Proc. WWW*. Hyderabad, India, 227–236.

Jie Bao, Jiao Tao, Deborah L. McGuinness, and Paul Smart. 2010. Context Representation for the Semantic Web. In *Proc, Web Science Conference*. Raleigh, USA.

Xin Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. 2010. SOLOMON: Seeking the Truth Via Copying Detection. *PVLDB* 3, 2 (2010), 1617–1620.

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. Truth Discovery and Copying Detection in a Dynamic World. *PVLDB* 2, 1 (2009), 562–573.

Xin Luna Dong and Divesh Srivastava. 2013. Big Data Integration. (Tutorial). In *Proc. ICDE*. Brisbane, Australia, 1245–1248.

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating Information from Disagreeing Views. In *Proc. WSDM*. New York, USA, 131–140.

François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. 2013. Fact Checking and Analyzing the Web. In *Proc. SIGMOD*. New York, USA, 997–1000.

V. Goranko and M. Otto. 2006. *Handbook of Modal Logic*. Elsevier. 255–325 pages.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical Extraction of Disaster-Relevant Information from Social Media. In *proc. WWW 2013*. Rio de Janeiro, Brazil, 1021–1024.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent Features of Rumor Propagation in Online Social Media. In *Proc. ICDM*. Dallas, USA, 1103 – 1108.

Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. 2012. Truth Finding on the Deep Web: Is the Problem Solved? *PVLDB* 6, 2 (2012), 97–108.

Fenglong Ma, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. 2015. FaitCrowd: Fine Grained Truth Discovery for Crowdsourced Data Aggregation. In *Proc. SIGKDD*.

Jeff Pasternack and Dan Roth. 2013. Latent Credibility Analysis. In *Proc. WWW*. Rio de Janeiro, Brazil, 1009–1020.

Krishnaprasad Thirunarayan, Pramod Anantharam, Cory Henson, and Amit Sheth. 2014. Comparative Trust Management with Applications: Bayesian Approaches Emphasis. *Future Generation Computer Systems* 31, 0 (2014), 182–199.

M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller. 2014. Continuous Data Cleaning. In *Proc. ICDE*. Chicago, USA, 244 – 255.

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2008. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE Trans. Knowl. Data Eng.* 20, 6 (2008), 796–808.

Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *PVLDB* 5, 6 (2012), 550–561.

Received November 2015; revised **** ****; accepted **** ****