Actionable Truth Discovery Challenges

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

In the Web, a massive amount of user-generated contents, public, open, or crowd-sourced information are available through various channels such as tweets, RSS, websites, DBs, multimedia-sharing platforms, social media and networks, etc. False information, rumors, and fake contents across multiple sources can be easily spread, making it hard to distinguish between what is true and what is not. Given such a large number of multi-channel information sources and the vast volume of conflicting data, ascertaining the veracity of the information available in the Web in a scalable and timely manner is extremely challenging and has only been preliminarily and partly addressed by existing work. In this paper, we present a set of research challenges related to truth discovery in the Web and we propose some potential solutions for verifying the veracity of online information. We also discuss how an integrative solution should benefit from a variety of technologies such as natural language processing, information extraction, data integration, probabilistic inference and data analytics, to help users estimate the reliability of various sources, check the truthfulness of their information, and turn this into actionable knowledge.

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 (January 2015), 8 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 media and social networks is growing in many domains for a variety of private as well as corporate usages. One of the fundamental difficulties is that data can be biased, noisy, outdated, incorrect, misleading and thus unreliable. Massive data coming from multiple sources amplifies this problem since conflicting information have to be "aligned", compared and check to estimate their veracity. For obvious reasons, 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[1]. Online fact-

[1]For example, the *Fog Computing* project from DARPA is a prototype developed in response to Wikileaks for automatically generating and distributing believable misinformation and then tracking access and attempted misuse of it - http://www.dtic.mil/dtic/tr/fulltext/u2/a552461.pdf

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

checkers (e.g., FactCheck[2], Snopes[3], PolitiFact[4], TruthorFiction[5] or OpenSecrets[6]) and humanitarian initiatives (e.g., AIDR[7] [Imran et al. 2013]) have gained unprecedented attention since they play an essential role in classifying and verifying manually (or semi-automatically) online information.

Beyond social media and computational journalism, the problem of truth discovery is also intellectually and technically interesting enough to have attracted a lot of prior studies, from both the artificial intelligence and the database communities, sometimes investigated 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 claimed by multiple sources, the ultimate goal of truth discovery is to label the claimed information items as true or false and compute the reliability of their sources, with the assumption that no *a priori* knowledge of the truthfulness of the individual assertions and sources is given. Relaxing this assumption has led to one major line of previous work which proposed and extended fact-finding models by incorporating prior knowledge either about the claimed 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 updating the trustworthiness of a source as a function of the belief in its claims, and then the belief score of each claim as a function of the trustworthiness of the sources asserting it [Yin et al. 2008]. In this line, several probabilistic models have been proposed to incorporate various aspects beyond source trustworthiness and claim belief, namely: the dependence between sources [Dong et al. 2010], the temporal dimension in discovering evolving truth [Dong et al. 2009], the difficulty of ascertaining the veracity of certain claims [Galland et al. 2010], the management of complex data structures such as collections of entities in a claim [Zhao et al. 2012] or linked data [Goasdoué et al. 2013]. However, current approaches are still limited in scope and also suffer from substantial drawbacks: limiting assumptions in the truth discovery models, opacity, complex model parametrization, and lack of scalability make them difficult to adapt or apply to the wide diversity of information available on the Web and for 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 conflicting and false information coming from various channels, in different languages, formats, and at different paces. Truth discovery systems have to be designed to help institutions and citizens with providing rigorous, scientific explanations and reports of their findings along with collected evidences. Although some preliminary work have addressed parts of this vision, e.g., [Dong and Srivastava 2013; Dong et al. 2010; Li et al. 2012], we argue that there is 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.

In this paper, we present a set of research chaleenges that together aim to effectively address the problem of truth discovery in real-world over-information scenarios. We

[2] www.factcheck.org/
[3] www.snopes.com/
[4] www.politifact.com/
[5] www.truthorfiction.com/
[6] www.opensecrets.org/
[7] http://aidr.qcri.org/

describe the challenges we face, propose our vision of a preliminary solution, and lay down a research agenda that goes beyond the Database community.

2. ACTIONABLE TRUTH DISCOVERY

2.1. An Illustrative Example

"The first casualty when war comes, is truth," said American Senator Hiram Johnson in 1917. For reporters, photographers, and cameramen captured the realities of wars and armed conflicts, sound and verifiable data is extremely difficult to gather and determining the complete picture of what occurred during a conflict is hard and perilous. Data is generally used effectively to answer a specific question (e.g., how many civilian casualties in Syria?) but it is often stretched to answer a broader question for which it was not suited, blurring the lines between "what was observable" and "what was true". In this context, sources of information can include victims or witness reports and online volunteered information (e.g., in English or Arabic) from personal and institutional Web pages, social and traditional media, NGOs', governmental and administrative records[8]. Seeking the truth is a tedious process manually done by war correspondents, United Nations experts and commissioners, Truth Commissions[9], human rights activists, journalists, and historians (e.g., WITNESS[10], StoryFul[11]). For this type of application, automating continuous truth discovery from the Web and social media networks to support decision-making and provide timely response to crisis situations goes far beyond data integrity checking, data quality, and data fusion.

While the challenges identified hereafter are by no means exhaustive, they are certainly important issues to address in priority for designing a truth discovery system.

2.2. Timely and Actionable Truth Discovery

Discovering truth from past events and historical data is certainly useful and results may be validated more easily since ground truth and full data sets already exist before the time of analysis. However, from a humanitarian perspective, actionable truth discovery from quasi real-time data could save lives. In this context, information extraction and triage as well as truth discovery computation need to be streamlined, prioritized, and adjusted to the degree of emergency and incompleteness of available information. This important usage-driven aspect goes beyond consistency checking and constraints. For example, information such as "The Chrysler building had an explosion and it affected the Empire State building" can be automatically verified: these buildings are about 2 km away and hence this information can not be true. However, more often what journalists/emergency responders need, is to verify information such as "The Chrysler building had an explosion" when there are very few primary sources claiming it initially. Timely and actionable truth discovery requires the prioritization and adjustment of information checking tasks specifically to the communities that will use the data.

2.3. 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 defined as the ability of the system to efficiently extract and map information: *i)* from various languages, *ii)* in various formats and

[8] e.g., http://www.correlatesofwar.org/
[9] http://www.truthcommission.org/
[10] http://www.witness.org
[11] http://storyful.com/

data structures, and *iii)* supported by various media and technologies. Traditional information extraction pipelines require several analysis stages ranging from text preprocessing to entity resolution and normalization. For each content, once entities, attributes, and relations are recognized and extracted, statements are identified, and sentences are classified (e.g., forecast or not), they have to be linked across multiple contents in different languages and eventually translated into a common language (e.g., English). A number of applications have been deployed[12] for information extraction, but most of them have used hand-crafted lists of terms and regular expressions, rather than corpus-trained approaches. They are usually specific to a single channel/application/language and cannot be considered "agile" enough for extracting, gathering and aligning information from various languages, formats or channels. Moreover, each stage of textual content analysis (from preprocessing, entity matching to classification) produces systematic and random errors that need to be considered in truth discovery computation. None of the very few attempts to couple these applications to truth discovery systems (e.g., [Goasdoué et al. 2013]) has considered the uncertainty from information extraction and processing affecting the truth discovery results. Tracing and estimating the errors of information extraction, formatting, and linking is one of the main challenges for automating truth discovery.

2.4. Incomplete and Biased Observations

Observation data may be incomplete and biased for various reasons: some information sources may not give all their data for security or privacy concerns; some sources have format limitations; when the sources do not explicitly provide temporal or spatial information, the time and location values are missing and need to be inferred. Typically, the Open World assumption holds since not all the real-world entities/events may have been observed and reported by the sources. Online sources have various levels of *a priori* knowledge on how the observation data have been collected. In many cases, opportunistic (or volunteered) data are *biased by observation effort* and the underlying observation method is unknown. It is thus important to explore how one can retrieve and quantify the observation effort from modeling the data provider's/observers' distribution. Data can be biased in many other ways. For example, data can be affected by a *selection bias* that is the extent to which a piece of information is claimed by a category of individuals who are likely to have access to and choose to use a certain type of media or channel[13]. Data may also suffer from the *observer's bias* (related to the individual's ability to accurately remember and describe events) and *disclosure bias* (related to witness's incentive (or disincentive) to include certain events or details). One challenge is to estimate the incompleteness and biases and take them into account in the truth discovery computation.

2.5. Decontextualization and Distortion

When a piece of information is extracted from its original content and channel, it may lose its context along with important "semantic markers" to understand *when, where, how, why,* and *for which purpose/audience* this particular piece of information has been produced, what it is supposed to mean in the absolute sense and also relatively to its particular context and information channel (e.g., a tweet in a thread). Information without context can be easily distorted and misinterpreted. For example, ascertaining

[12]For example, Europe Media Monitor's News Explorer (http://emm.newsexplorer.eu) gathers news, clusters related news stories, and extracts names, locations, general person-person relations, and event types; Open-Calais from Thomson Reuters (http://www.opencalais.com) extracts a range of entity, relation, and event types from general and business news.

[13]e.g., social media users are usually younger, technologically savvy, motivated individuals versus individuals who are older, more remote, lacking access or ability or motivation to use technology and whose stories/observations are likely missing from this channel.

the veracity of a controversial information that has been re-tweeted many times in different contexts by various users with different emotions and motivations is difficult without the description of these contexts. Moreover, the pace has to be considered when information updates from Twitter, Facebook, and Wikipedia have to be "aligned" with considering the source granularity (*i.e.*, a source can be an individual, an organization, a company, a government, etc.). Much research work has studied the formalization of contexts and rich context representations have been proposed [Bao et al. 2010]. However, current information extraction methods constrain 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.

2.6. Source Expertise and Information Correlation

In many domains, some sources are more authoritative or specialized, providing accurate information for a small subset of real-world objects, while other sources are generalist with a wider coverage, providing a huge amount of information, some of which may be out-of-date or imprecise. Determining truth between such sources based only on majority voting and similarity of values can easily lead to erroneous conclusions. Hence, any approach for determining truth needs to consider the possibility that sources have different levels of expertise, various updating policies, and they provide different coverages of information over time. Moreover, similar values across many sources may be due to the likelihood that some information items are either very specific (rare) or very general (popular) and/or sometimes highly correlated; this does not necessarily mean that the sources are dependent but information correlation and distribution have to be analyzed before truth discovery computation.

2.7. Limitations of Truth Discovery Models

Most truth discovery models suffer from limiting assumptions on the claims, referred real-world objects, sources, and truth discovery results. For example, one important assumption on the claims is that the assertions made by the sources (and whose veracity is unknown) are organized into disjoint mutual exclusion sets and exactly one of the claims in each mutual exclusion set is assumed to be true. Claims are also assumed to be positive[14] and independent, as well as real-world objects they refer to. Concerning the sources: a source is supposed to contribute uniformly to all the claims it expresses. Sources are assumed to be independent (except in [Dong et al. 2010]), to make their assertions independently and do not provide conflicting claims. Moreover, the probability a source asserts a claim is independent of the truth of the claim and there is explicit correspondence between the sources and the claims (*i.e.,* the models do not consider cases such that "S_1 claims that S_2 claims A"). Finally, each claim is assumed to be either true or false. No uncertainty, gradual result or ranking is supported by current models for labeling the truthfulness of claims. Relaxing these assumptions constitutes a challenging research project that could be overreached by formally defining the semantics of truth discovery and making it operational in realistic cases.

2.8. Multimodal Truth Discovery

** Laure to complete **

2.9. Dynamics of Truth Discovery

** Laure to complete **

[14] *i.e.,* Cases such as "S claims that A is false" or "S does not claim A is true" are not considered.

3. TOWARDS AN INTEGRATIVE SOLUTION

3.1. Semantics for Truth Discovery

Truth discovery from multi-source data relies on the Open World Assumption with overlapping and conflicting data. Designing an appropriate data model for truth discovery requires a fine trade-off between tractability and expressiveness. The vision of DAFNA is to give a concrete and unified semantics to truth discovery problem. One principled solution –which can: *i)* relax most limiting assumptions of previous models, *ii)* formally unify existing approaches, and *iii)* support inference– is to define formal semantics of truth discovery systems in modal logics through axioms and canonical *Kripke* structures [Goranko and Otto 2006]. However, reasoning in modal logics can quickly become intractable [Gottlob 1992]. This applies, in particular, to information items that include possibility in addition to certainty and impossibility (positive or negative claims). In this work, we have chosen to define the semantics of a truth discovery system such as it supports fragments of modal logics to allow negations on claims (which is sufficient to express conflicts) and we introduce a particular *Kripke* structure that allows claim verification procedure and powerful inferring functionalities (for chaining evidences). Few work in databases have considered modal logics before: Calvanese *et al.* [Calvanese et al. 2008] use the modal logic $K45_n^A$ to allow mappings between peers in the presence of conflicts. Gatterbauer et al. [Gatterbauer et al. 2009] introduce the notion of *belief database* to manage conflicting annotations. In our opinion, formal semantics and reasoning based on multi-agent epistemic logic for data integration deserves much more attention than it has received so far from the Database community. In DAFNA, our vision is to leverage modal logics for defining in a principled way the semantics of truth discovery and reasoning about multi-source conflicting claims. This constitutes a novel, alternative solution to current approaches in data fusion and integration.

3.2. Continuous Truth Discovery, Inference, and Belief Revision

None of the existing truth discovery models continuously incorporate new evidences, new claims, contextual metadata, prior beliefs, and background knowledge in their computation. They rely on *ad hoc* parametrization and do not provide revision, *what-if* analysis or explanation of their results. The vision of DAFNA is to automate claim verification continuously and adaptively for dynamic environments (e.g., for emergency response or humanitarian scenarios such as the one illustrated in Section 2). The system continuously searches and finds supporting and opposing evidences, as well as contextual meta-information from the Web of Data (B1); it adapts and revises its knowledge and belief of the sources/claims and recompute the truthfulness of the claims (B4). DAFNA is in-line with recent work on continuous cleaning [Volkovs et al. 2014] and belongs to the new family of dynamic data quality techniques in database systems. But it is intended to go beyond dynamic data consistency checking since the system incorporates contexts, evidences, and error estimates in addition to the set of claims; it supports evidence-based reasoning and inference with uncertainty and bias estimations and handles, in a scalable way, the complete information processing pipeline and truth discovery inference.

3.3. Diagnostics and Error Propagation Monitoring

There is a large body of work on managing uncertain and incomplete information [Green and Tannen 2006; Agrawal et al. 2010]. Our work shares a similar motivation for information that is not certain. However, we do not only measure, track, evaluate uncertainty (and biases) of user-generated contents, but DAFNA system also monitors uncertainty generated by the system itself in the information processing. Errors are

estimated at each stage of the truth discovery process and incorporated into the final results' computation by the key module (B5) *Diagnostics and Error Propagation* in Figure 1. DAFNA infrastructure supports aggregation of system state and allows the user to specify *ad hoc* monitoring tasks by using lightweight Event-Condition-Action (ECA) rules applied to each module: information extraction (B1), data preprocessing and alignment (B2), data and source quality profiling (B3), and truth discovery and revision (B4). In that sense, DAFNA is a diagnostic system that is able to analyze information on its performance and accuracy at each stage and can incorporate this information in its results, either for validation, error estimation or for supporting and documenting clear explanation and visualization of the results provided to the users by (F1), (F2), and (F3) front-end modules.

4. CONCLUDING REMARKS

The Web of Data has accelerated the rate at which information is produced and disseminated through various channels and medias, but has also eased the ability to spread false information. Whereas previous work on fact-checking from multiple sources focused on one single channel, one language, limiting assumptions, with important scalability issues, current approaches ignore information context, often assume source independence, and fully accurate information extraction and entity linking. In this paper, we have identified the main issues that are of particular interest for designing actionable truth discovery systems, discussed preliminary ideas on how we can overcome these challenges, and enumerated the key components in the design of an integrative solution. We have outlined several research opportunities that arise from taking a more principled view of actionable truth discover. We expect research along this line can help users better understand and ascertain online information veracity in the Web of Data. Actionable truth discovery gives clear opportunities to the Database community for joint research with natural language processing, logics, and social computing communities, not only to reflect the entire process of truth discovery in the Web of Data, but also to have a technological and human impact in providing operational data veracity management systems that effectively support the fourth "V" (Veracity) of Big Data.

REFERENCES

Parag Agrawal, Anish Das Sarma, Jeffrey D. Ullman, and Jennifer Widom. 2010. Foundations of Uncertain-Data Integration. *PVLDB* 3, 1 (2010), 1080–1090.

Raju Balakrishnan and Subbarao Kambhampati. 2011. SourceRank: Relevance and Trust Assessment for Deep Web Sources based on Inter-source Agreement. In *The 20th International World Wide Web Conference (WWW 2011)*. ACM, Hyderabad, India, 227–236.

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

Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. 2008. Inconsistency tolerance in P2P data integration: An epistemic logic approach. *Inf. Syst.* 33, 4-5 (2008), 360–384.

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 *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, Brisbane, Australia, 1245–1248.

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. 2010. Corroborating Information from Disagreeing Views. In *Proceedings of the third ACM international conference on Web search and data mining (WSDM)*. ACM, New York, USA, 131–140.

Wolfgang Gatterbauer, Magdalena Balazinska, Nodira Khoussainova, and Dan Suciu. 2009. Believe It or Not: Adding Belief Annotations to Databases. *PVLDB* 2, 1 (2009), 1–12.

François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, and Stamatis Zampetakis. 2013. Fact Checking and Analyzing the Web. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD'13)*. ACM, New York, USA, 997–1000.

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

Georg Gottlob. 1992. Complexity Results for Nonmonotonic Logics. *J. Log. Comput.* 2, 3 (1992), 397–425.

Todd J. Green and Val Tannen. 2006. Models for Incomplete and Probabilistic Information. *IEEE Data Eng. Bull.* 29, 1 (2006), 17–24.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical Extraction of Disaster-Relevant Information from Social Media. In *The 22nd International World Wide Web Conference (WWW 2013)*. ACM, 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 *2013 IEEE 13th International Conference on Data Mining (ICDM)*. IEEE, 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.

Jeff Pasternack and Dan Roth. 2013. Latent Credibility Analysis. In *The 22nd International World Wide Web Conference (WWW 2013)*. ACM, 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. DOI:http://dx.doi.org/10.1016/j.future.2013.05.006

M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller. 2014. Continuous Data Cleaning. In *2014 IEEE 30th International Conference on Data Engineering (ICDE)*. IEEE, 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 **** ****