The uncertain web: concepts, challenges, and current solutions

Djamal Benslimane, Quan Z. Sheng, Mahmoud Baarhamgi, Henri Prade

▶ **To cite this version:**

Djamal Benslimane, Quan Z. Sheng, Mahmoud Baarhamgi, Henri Prade. The uncertain web: concepts, challenges, and current solutions. ACM Transactions on Internet Technology, 2016, 16 (1), pp.1-6. 10.1145/2847252 . hal-03514798

HAL Id: hal-03514798
https://hal.science/hal-03514798

Submitted on 6 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive Toulouse Archive Ouverte

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

This is an author's version published in: http://oatao.univ-toulouse.fr/24810

Official URL: http://dx.doi.org/10.1145/2847252

To cite this version: Benslimane, Djamal and Sheng, Quan Z. and Baarhamgi, Mahmoud and Prade, Henri *The uncertain web: concepts, challenges, and current solutions.* (2016) ACM Transactions on Internet Technology, 16 (1). 1-6. ISSN 1557-6051

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr

The Uncertain Web: Concepts, Challenges, and Current Solutions

DJAMAL BENSLIMANE, Claude Bernard University Lyon 1
QUAN Z. SHENG, The University of Adelaide
MAHMOUD BARHAMGI, The Open University & Claude Bernard University Lyon 1
HENRI PRADE, Paul Sabatier University & University Technology Sydney

1. INTRODUCTION

Uncertainty, incompleteness, and imprecision are common characteristics of the data and knowledge that we daily deal with in a wide range of domains and applications. Uncertainty management has been extensively studied in databases, artificial intelligence, and operations research for several decades. In the area of Internet technologies, uncertainty handling is an issue in e-commerce, social and sensor networks, scien-tific data production and exploration, objects tracking, data integration, location-based services, open linked data, and, recently, the Web of Things. For example, in the e-commerce domain alone, some recent studies [Soliman et al. 2010] have shown that approximately 65% of products (e.g., properties, cars, and so on) that one would find on typical e-commerce sites (e.g., apartments.com, carpages.ca) are associated with some uncertainty in their basic information (e.g., prices, locations, and descriptions).

A common factor to most of these domains is the growing reliance on the World Wide Web (WWW) as an integrated platform for collecting, storing, processing, managing, querying, and servicing this uncertain data to users. The WWW has undoubtedly become an immense sea of interconnected uncertain data sources and uncertain services [Wang et al. 2015]. Exploiting these uncertain data sources and services to their full potential raises important research challenges that relate to the different phases of their lifecycle, including:

- *Data and Service Creation:* As the uncertainty of a piece of information or of a service is important for understanding its semantics and using it correctly, it should be measured, quantified (or qualified), and associated with that information or service.

Authors' addresses: D. Benslimane, LIRIS Lab, Claude Bernard University Lyon-1, 69622 VILLEURBANNE CEDEX, France; email: djamal.benslimane@univ-lyon1.fr; Q. Z. Sheng, School of Computer Science, The University of Adelaide, SA 5005, Australia; email: michael.sheng@adelaide.edu.au; M. Barhamgi, The Open University, Walton Hall, Milton Keynes, Buckinghamshire MK7 6AA, United Kingdom; email: mahmoud. barhamgi@univlyon1.fr; H. Prade, IRIT, Paul Sabatier University, 31062 Toulouse cedex 9, France, & QCIS, University of Technology Sydney, Ultimo NSW 2007, Australia; email: prade@irit.fr.

- *Data and Service Representation:* Modeling and representing the semantics of Web data sources and services is the first step toward the automation of their different related activities such as data and service querying, selection, integration, ranking, and composition. As the uncertainly of data and services could impact these activities, it should be considered as an integral part of data and service descriptions. Current Web modeling languages and standards (e.g., RDF, OWL, micro formats, WSDL, REST, and OWL-S) should be extended and used to represent the uncertainty of data and services.
- *Data and Service Consumption:* An uncertain data item could be consumed in isolation or combined to other data items from other data services to derive value-added information. Efficient techniques and models are required to efficiently query uncertain data pieces and integrate them to answer user's queries. Such techniques should also aggregate the uncertainties of integrated data pieces to allow the user to correctly interpret and use the query's result.

In this article, we review some of the most important challenges related to uncertain Web data and services. We also present, in Section 3, some of the latest developments in this research field. Specifically, we report six research papers addressing different aspects of the uncertain Web.

2. UNCERTAIN WEB RESEARCH CHALLENGES

Uncertainty raises a good number of research challenges in the context of Web computing. In this section, we discuss some of the most important ones.

2.1. Modeling of Uncertain Web Data and Services

Different models have been proposed for representing data uncertainty in uncertain relational databases. The *Possible Worlds Semantics* [Suciu et al. 2011] is an important concept for understanding the models of uncertain data. In the possible world semantics, data uncertainty is captured by viewing a relational table as a set of possible instances that correspond to the different possible instantiations of the uncertain data items. Many uncertainty models, for example, Abiteboul et al. [1987] and Imielinski et al. [1984], adopt the possible worlds semantics, where a probabilistic database D is viewed as a set of possible instances (worlds) $\{PW_1 \ldots PW_n\}$. The possible worlds space represents an enumeration of all possible views of the database resulting from the uncertainty or incompleteness in the underlying data. One of the important models that adopt possible worlds semantics to capture uncertainty and incompleteness in attribute values is the c-tables model [Imielinski et al. 1984]. C-tables are relational tables whose attributes are represented using variables, and each tuple is associated with a Boolean condition on the attribute variables. A tuple belongs to the database if and only if its associated condition is satisfied. In addition to probabilistic models, possibility theory is another noteworthy example of a setting based on possible world semantics [Bosc and Pivert 2010].

Research efforts are needed to adapt these uncertainty models to data and service representation models and standards of the Web including, RDF, RDFS, OWL, XML, Microdata, Linked data, WSDL, REST, OWL-s, and SA-WASDL. This prepares the ground for automating the different activities related to Web data and services, including search, querying, integration, composition, interpretation, and exploitation. In addition, it is worthwhile to mention the works in probabilistic (or possibilistic) description logics, such as Lukasiewicz and Straccia [2009] and Benferhat and Bouraoui [2013].

2.2. Integration of Uncertain Data Sources and Services

While there has been considerable past work studying data integration and uncertain data in isolation, only few works have addressed the data integration problem when the sources to be integrated are uncertain. Still, artificial intelligence researches in information fusion or on ontology alignments might be sources of cross-enrichment for ontology matching [Gal and Shvaiko 2009]. In Dong et al. [2009], a Local-as-View-like data integration system was proposed for uncertain data sources. However, this work has addressed only the issues of creating the probabilistic mappings between the mediated schema and the data sources' schemas, as well as query transformation based on these probabilistic mappings. Along the same lines, in [Magnani and Montesi 2010], the authors survey the different approaches (and the used formalisms) for the construction of probabilistic mappings and the corresponding query transformation mechanisms. However, all of these works have addressed the uncertainty at the schema level only; that is, the uncertainty resides in the way sources can be mapped to the mediated schema.

More research efforts are needed to address the uncertainty at the data level, that is, when the data inside the sources to be integrated are themselves uncertain. New models, techniques, and solutions are needed to aggregate the uncertainties of data sources involved in answering a query and computing the uncertainty of the query's result. Similarly, Web services composition techniques should be extended to compute the uncertainty of a composition's output when the composed services are uncertain.

2.3. Efficient Query Evaluation Techniques

Two mainstream approaches have been proposed for query evaluation over uncertain data, the *intentional* and *extensional* approaches [Dalvi and Suciu 2007]. In the intentional approach, the probability of each tuple is represented as an event-based probabilistic formula. When a query is processed, the probabilistic formulas of intermediate tuples and the final result are computed. However, the intentional approach incurs a prohibitive computation cost and is thus impractical for real applications [Re and Suciu 2007]. In contrast, the extensional approach is quite efficient and is based on rewriting the query plan using the probabilistic relational algebra. However, in this approach, not all of the possible query plans compute the correct probabilities; the ones that do are referred to in the literature as *safe plans*. Finding the most efficient query plans that still correctly compute the probabilities of a query result is an active research challenge. In some Web applications, such as Web objects ranking, the most important task is to efficiently rank objects (based on their probabilities) rather than to know their exact probabilities. Therefore, an unsafe, but efficient, query plan that would compute approximate probabilities (but precise enough for the ranking purpose), would be preferred sometimes over a safe, but inefficient, query plan.

Therefore, more research efforts are needed to study the problem of finding the safe plans of a query and evaluating their computation costs, and to study and quantify the probability error bounds that could be produced by an unsafe query plan. This will make it possible to rank the possible plans based on their efficiency and probability error bounds, and to choose the best query plan that better meets the requirements of the considered application. It has been recently shown that a data complexity identical to the one of the classical database case can be preserved in the possibilistic setting if we are only interested in the answers that are certain to some extent (and not in those that are just possible) thanks to the existence of a strong representation system result [Pivert and Prade 2015]. Possibilistic keys, which are keys associated with a certainty level that says to which tuples the key applies, can be useful for cleaning dirty data [Koehler et al. 2014].

2.4. Ranking of Uncertain Data and Services

Queries over uncertain Web data sources often return an overwhelming number of results (e.g., data tuples), leading data consumers to miss the ones that are most relevant to their needs. Top-k queries are a common approach to report the best k answers (of a query) based on matching the processed data tuples to users' preferences. In the context of uncertain data sources and services, data items should be ranked based not only on their matching degrees with users' preferences, but also on their uncertainties, i.e., data items' scores and uncertainties interplay to decide the top-k output data. The interaction between data uncertainty and the "top-k" gives rise to different possible interpretations of uncertain top-k queries: (1) the top-k tuples in the most probable world; (2) the most probable top-k tuples that belong to valid possible world(s); and (3) the set of most probable top-i-th tuples across all possible worlds, where $i = 1 \ldots$k. However, if the preferences are quantified, an expected value criterion may also be used for ranking the answers, both in the probabilistic and in possibilistic settings.

More research efforts are needed to devise new ranking techniques for uncertain data and services that consider the uncertainty as an important ranking dimension, along with the other QoS-based dimensions.

3. IN THIS SPECIAL ISSUE

Six articles have been selected in this special issue after several rounds of rigorous review by the guest editors and the invited reviewers.

In "Towards Anomalous Diffusion Sources Detection in a Large Network," Peng Zhang, Jing He, Guodong Long, Guangyan Huang and Chengqi Zhang address the problem of malicious information diffusion in large networks. Malicious information, such as rumors, virus, and spam, are everywhere and often cause severe damage to our society. The authors develop an efficient method to detect anomalous diffusion sources, thus protect networks from security and privacy attacks. The novelty of the proposed method is that it needs only a small set of network nodes or detectors, as opposed to the existing works and methods that are based on the assumption that network snapshots reflecting information diffusion can be obtained continuously, an assumption that would require deploying detectors on all of the network nodes. The authors propose a new regression learning model that can detect anomalous diffusion sources by jointly solving five challenges: (1) unknown number of source nodes, (2) few activated detectors, (3) unknown initial propagation time, (4) uncertain propagation path, and (5) uncertain propagation time delay. The proposed model is theoretically and empirically analyzed and tested; the obtained results prove its efficiency and good performance compared to existing solutions.

In "Quality-Based Online Data Reconciliation," Asma Abboura, Soror Sahri, Latifa Baba-Hamed, Mourad Ouziri, and Salima Benbernou address the problem of duplicates detection while answering user queries over databases with redundant information about same world entities. The proposed approach detects the duplicates at the query answering time, as opposed to the existing works that address the problem by eliminating the duplicates from underlying data sources. The approach is based on the techniques of Matching Dependencies (MDs) to detect duplicates through the concept of Data Reconciliation Rules (DRR), and the Conditional Function Dependencies (CFDs) to assess the quality of different attribute values. In addition, the approach relies on a duplicate reconciliation index (DRI), constructed based on clusters of duplicates detected by a set of DRRs, to speed up the online data reconciliation process. The reported experimental results show the efficiency and effectiveness of the proposed solution.

In "Supervised Anomaly Detection in Uncertain Pseudo-Periodic Data Streams," Jiangang Ma, Le Sun, Hua Wang, Yanchun Zhang, and Uwe Aickelin address the

issue of anomaly detection in uncertain data streams, which is an important problem in a wide range of Web applications. Anomaly detection is the process of finding abnormal behaviors from given data streams. The uncertainty in data streams makes anomaly detection from sensor data streams a very challenging issue. For example, in an ECG data stream, if a sensor error is classified as abnormal heartbeat signals, it may cause a serious misdiagnosis. The authors propose an integrated framework to support anomaly detection in uncertain data streams. The proposed framework defines an efficient uncertainty preprocessing procedure to identify and eliminate uncertainties in data streams. It also defines a set of efficient pattern recognition and feature extraction techniques. The framework exploits mature classification methods for anomaly detection in the corrected data stream. The authors have validated their techniques using a real ECG dataset. The reported experimental results show that the proposed techniques outperform previous approaches in terms of accuracy in anomaly detection.

In "Using an Epidemiological Approach to Maximize Data Survival in the Internet of Things," Abdallah Makhoul, Christophe Guyeux, Mourad Hakem, and Jacques M. Bahi focus on the data survivability issue in large-scale and self-coordinating IoT (Internet of Things)-based environments in the presence of uncertain situations. Examples of such uncertain situations include the disruption of a critical monitoring infrastructure when a portion of it becomes unreachable, data collection infrastructures in urban disaster areas, which can be used to predict natural hazards and save lives and which could be affected by nodes failures or attacks. Such uncertain situations may lead to a significant loss of data. The authors propose a new model inspired by epidemic and disease propagation models to save the data in the network and make it survive after attacks. The proposed model, dubbed the e-Epidemic SIR (Susceptible-Infectious-Recovered) model, takes into account the dynamicity of the network topology and the energy constraints of its nodes. The authors report a set of simulations that showcase the efficacy of their solution.

In "Constructing Maintainable Semantic Relation Network from Ambiguous Concepts in Web Content," Kenneth Wai-Ting Leung, Di Jiang, Dik Lun Lee, and Wilfred Ng propose a new methodology to construct a Concept Relation Network (CRN) that can be used to represent the different possible interpretations of a concept and measure their degrees of uncertainty. The proposed methodology relies on the use of well-known Web search engines, such as Google and Yahoo, to extract the different possible interpretations of a concept and to characterize their probabilities in preparation for their inclusion into the constructed CRN. The uncertainty of a concept is defined using the entropy notion from the information theory, which is computed based on its number of possible interpretations, that is, the number of related concepts in the CRN. The constructed CRN is useful for many Web applications, including social networks, recommendation systems, and semantic based search engines. The article has exploited the constructed CRN in two particular domains: semantic information retrieval and Web analytics. The obtained experimental results show that CRN can enhance these applications by considering the heterogeneous and polysemous nature of Web content.

In "sCARE: Reputation Estimation for Uncertain Web Services," Zaki Malik, Brahim Medjahed, and Abdelmounaam Rezgui propose a trust-estimation approach, dubbed sCARE (Statistical Cloud-Assisted Reputation Estimation), for service-oriented environments in uncertain situations. The proposed approach consolidates and integrates the uncertain and fuzzy ratings submitted by different service consumers to provide a unified and holistic trust assessment of a given service provider. It relies on a statistical model to represent the uncertainty of service ratings that can be used to compute the credibility of raters as well as the trust of service providers.

4. CONCLUSION

Uncertainty is one of the most common characteristics of the data in a wide range of applications, particularly the services over the World Wide Web. In dealing with such uncertain data needs, there are many challenges to overcome that have been only partially addressed so far. The articles included in this special issue cover several important topics and present some of the key directions in this key area of research and development. We hope that the set of selected articles provides the community with a better understanding of the current directions and areas to focus on in the future, and inspires your own work.

ACKNOWLEDGMENTS

We thank all the authors for considering this special issue as an outlet to publish their research results in the area of context-aware Web services. We also would like to thank the referees who provided very useful and thoughtful feedback to the authors. Finally, we express our gratitude to the Editor-in-Chief, Professor Munindar P. Singh, for his kind support, advice, and encouragement throughout the preparation of this special issue.

REFERENCES

Serge Abiteboul, Paris Kanellakis, and Gosta Grahne. 1987. On the representation and querying of sets of possible worlds. *SIGMOD Record* 16, 3, 34–48.

Salem Benferhat and Zied Bouraoui. 2013. Possibilistic DL-Lite. In *Proceedings of the 7th International Conference on Scalable Uncertainty Management (SUM'13)*, Weiru Liu, V. S. Subrahmanian, and J. Wijsen (Eds.), Lecture Notes in Computer Science, Vol. 8078, Springer, Berlin, 346–359.

Patrick Bosc and Olivier Pivert. 2010. Modeling and querying uncertain relational databases: A survey of approaches based on the possible worlds semantics. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 18, 5, 565–603.

Nilesh Dalvi and Dan Suciu. 2007. Efficient query evaluation on probabilistic databases. *The VLDB Journal* 16, 4, 523–544.

Xin Luna Dong, Alon Y. Halevy, and Cong Yu. 2009. Data integration with uncertainty. *The VLDB Journal* 18, 2, 469–500.

Avigdor Gal and Pavel Shvaiko. 2009. Advances in ontology matching. In: T. S. Dillon, E. Chang, R. Meersman, K. P. Sycara (Eds.), *Advances in Web Semantics I – Ontologies, Web Services and Applied Semantic Web*. Lecture Notes in Computer Science, Vol. 4891, Springer, Berlin, 176–198.

Tomasz Imielinski and Witold Lipski, Jr. 1984. Incomplete information in relational databases. *Journal of the ACM* 31, 4, 761–791.

Henning Köhler, Uwe Leck, Sebastian Link, and Henri Prade. 2014. Logical foundations of possibilistic keys. *Proceedings of the 14th European Conference on Logics in Artificial Intelligence (JELIA'14)*, E. Fermé and J. Leite (Eds.). Lecture Notes in Computer Science, Vol. 8761, Springer, Berlin, 181–195.

Thomas Lukasiewicz and Umberto Straccia. 2009. Description logic programs under probabilistic uncertainty and fuzzy vagueness. *International Journal of Approximate Reasoning* 50, 6, 837–853.

Matteo Magnani and Danilo Montesi. 2010. A survey on uncertainty management in data integration. *Journal of Data and Information Quality* 2, 1.

Olivier Pivert and Henri Prade. 2015. A certainty-based model for uncertain databases. *IEEE Transactions on Fuzzy Systems* 23, 4, 1181–1196.

Christopher Re and Dan Suciu. 2007. Management of data with uncertainties. In *CIKM*. 3–8.

Mohamed A. Soliman, Mina Saleeb, and Ihab F. Ilyas. 2010. MashRank: Towards uncertainty-aware and rank-aware mashups. In *ICDE 2010*. 1137–1140.

Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. 2011. Probabilistic databases. In *Synthesis Lectures on Data Management*, Meral Özsoyoğlu (Ed.). Morgan & Claypool, San Francisco, CA.

Xianzhi Wang, Quan Z. Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, and Xue Li. 2015. An integrated Bayesian approach for effective multi-truth discovery. In *CIKM 2015*.