Uncertainty over Structured and Intensional Data

Antoine Amarilli

under the supervision of Pierre Senellart, Télécom ParisTech, Institut Mines–Télécom

1 Proposal

The World Wide Web contains vast quantities of information of an heterogeneous nature available to automated agents: trillions of Web pages, hundreds of millions of social messages per day on websites such as Twitter, hundreds of knowledge bases in the open linked data cloud with dozens of billions of semantic facts, some of them implying more facts through semantic reasoning rules.

Traditional data management techniques are not suitable to integrate such data, answer complex queries over it, and mine it, for multiple reasons. First, the data is *uncertain*, namely, it is incomplete and partially wrong, so that all operations over the data need to keep track of its provenance lineage or of its probabilistic annotations. Second, the data may be *structured* in different ways (RDF triples, social networks, XML documents, relational data, views with aggregates) that need to be integrated. Third, access to the data is *intensional*, namely, we cannot assume that all relevant information has been collected in a centralized data store, and we must instead access the various data sources sparingly according to our needs, and account for the cost of such accesses on multiple incomparable axes.

We propose to study the interaction of uncertainty, intensionality, and structure, develop the foundations of data management following those three desiderata, and investigate its applicability to knowledge acquisition from the Web.

2 Example Application

To better illustrate, let us take the example application of *mobility in smart cities*, i.e., transportation options, travel habits, traffic, etc., in and around a city; resources mentioned in the previous section can be used to collect and enrich data related to this application. In addition, in such a setting, domain-specific resources, not necessarily public, such as street cameras, red light sensors, air monitoring systems, etc., contribute to the available data.

Consider now a *knowledge acquisition need* from a user (say, a transport engineer) on this mass of data. It can be a simple query expressed in a classical query language (e.g., "What is the optimal way to go from this place to that place at a given time of day?"), a kind of pattern to mine from the data ("Find an association rule of the form $X \Rightarrow Y$ that holds among people commuting to this district"), or some higher-level business intelligence query ("Find anything interesting about the maritime traffic in the Singapore Strait in the past 24 hours").

Focus on one specific question. For example, a civil engineer wants to know the total traffic through a given road on an arbitrary day, in order to plan a renovation of that road. There are many ways to accomplish such a task:

- Use a computer vision program to analyze the street camera feeds and automatically extract each passing vehicle;
- Ask crowd workers to perform the same analysis;

- Do the same, but only a fraction of the day, and extrapolate the results;
- Use traffic data from Bing Maps API, correlated with external data about road characteristics;
- Send a team of expert traffic specialists to survey the road;
- etc.

Each of these ways, and every combination thereof, has a cost in terms of manpower, budget, processing time, bandwidth (the data is *intensional*), has a precision both as a prior and as a posterior after using the services (the data is *uncertain*), and uses intermediate data such as images, aggregated views, Web service APIs, of very different, sometimes complex, structures (the data is *heterogeneously structured*). The objective is to obtain an optimal solution, for instance up to an approximation tolerance. This example is fairly simple, but one should keep in mind that determining the traffic on a road may be just one component of a more complex information need, such as route planning.

3 Challenges

3.1 Structure

Structured data sources on the Web are extremely diverse. Some are dumps of relational databases (matching the traditional focus of database theory), or Web forms that can be used to query such databases (and can be seen as relational databases up to wrapper induction, etc.). Others are tree-shaped XML documents, or HTML documents (Web pages). Yet others are graph-shaped: the links between a collection of Web pages, social networking data, semantic graphs, etc. Such structure must be leveraged, and the different models must be integrated, when using the Web as a source of knowledge.

Structure may also intervene in more complex ways. For instance, if we decide to acquire data through a crowdsourcing platform, the structure of the data that we obtain depends on the queries that we posed, the exact form of which we are free to define. If we decide to use Web sources with aggregate information (such as the number of reported crimes per year in different cities), we can also decide to interpret them *structurally* as complex materialized views of the data of interest.

3.2 Uncertainty

An uncertain document is a possibly infinite set of possible states of the world, optionally weighted according to a probability distribution, which is represented concisely as a finite document in a certain language. The exact representations used depend on the structure of the underlying data: probabilistic relational databases [SORK11] store uncertain relational data and are designed to answer relational queries efficiently. For tree-like data such as XML, numerous probabilistic frameworks exist [KS13].

The different ways to represent uncertain data can be compared according to their expressiveness (which probability distributions can be represented by a given model? how are the representations computed?), their size (how large is the representation of a distribution? how far can a representation be compressed or approximated?), and their efficiency (what is the complexity of computing such a model from examples? of answering queries over such models?). The performance of the various models over these criteria and the relations between those models are not yet adequately understood.

In the context of the Web, uncertainty can be provided directly at the source level (either explicitly, or implicitly in the case of vague or incomplete information), or it can arise from collections of conflicting untrustworthy facts, or as the result of inherently uncertain processes (information extraction, named entity recognition, crowdsourcing, etc.). In contrast with the *local* management

of uncertainty where only the most probable hypothesis is retained, it is necessary to maintain uncertainty *globally*, throughout the entire process, up to and including query results.

3.3 Intensionality

When we consider a complex query, or a knowledge acquisition need from a user, that we wish to evaluate over Web sources, we cannot perform this using the usual *extensional* approach of centralizing all the data which may be of interest and running the query over it. Indeed, such data would be too large and too costly to gather. We must instead use the data sources directly, through the diverse *access mechanisms* that they provide, and access data sparingly because of its inherent heterogeneous *costs* and hard restrictions (bandwidth, time, policy rate limitations, etc.).

In this context, under expressive access pattern languages, it is already complex to determine if an access is *relevant* for a high-level query [BGS11]. It is even more challenging to devise *execution plans* to evaluate complex queries through the permitted accesses. This question has also been studied in the context of crowdsourcing, under less expressive access patterns, seen as the problem of deciding which crowd query to pose next based on the existing answers. This decision, of course, relies on the management of uncertainty to represent our current knowledge about the world and estimate its accuracy.

Another relevant use case for intensionality is to represent implicitly the consequences of facts through certain inference rules, such as those that are provided in ontologies over semantic Web sources; of course, uncertainty needs to be taken into account when applying possibly uncertain rules to uncertain facts, one important difficulty being that there may be multiple non-independent ways to derive one fact.

4 Status

Our PhD work has started in September 2013. In this section, we summarize the preliminary work that we have undertaken towards the above goals, and illustrate specific possible ways in which we intend further research on these different topics, the list being non-exhaustive.

- In collaboration with Daniel Deutch from Tel Aviv University and M. Lamine Ba at Télécom ParisTech, we have investigated the question of maintaining *provenance information* over relational data with additional structure information, a generalization of the problem of maintaining probabilistic annotations to represent uncertainty throughout query evaluation. We have focused on data that is structured by incomplete orderings [ABDS13]. We plan to pursue our investigation of provenance annotation frameworks on rich structures, including XML trees, and to develop practical implementation of such frameworks.

- We have studied the *open-world query answering* problem of running complex queries on the intensional facts resulting from the completion of a relational instance by rules, and extended decidability results for this problem to expressive constraint languages [Ama13], under the supervision of Michael Benedikt from the University of Oxford. We intend to extend this approach to uncertain rules, in collaboration with Pierre Bourhis, CNRS researcher at the University of Lille. We also intend to study the practical usefulness of such schemes for query answering over knowledge bases using fuzzy deduction rules inferred from the data, for instance over the YAGO ontology maintained by Fabian Suchanek at Télécom ParisTech.

- We have achieved a better understanding of the interaction between probability and structure by connecting the probabilistic relational and XML models [AS13], and studied the tradeoff between expressiveness and computational complexity in the membership problem for possible worlds of probabilistic XML documents in various frameworks [Ama14]. We plan to invest

more effort in studying the foundations of probabilistic data management, for instance by characterizing sufficient conditions on the structure of uncertain relational data to ensure the tractability of query evaluation.

- In collaboration with Yael Amsterdamer and Tova Milo from Tel Aviv University, we have developed methods to choose accesses to intensional resources in the context of crowdsourcing. We have focused on the data mining task of frequent itemset identification using a domain ontology, and studied its computational and crowd complexity [AAM14]. We pursue this line of work by studying how uncertainty can be managed in such situations.

Supervision and Environment

This PhD project is supervised by Pierre Senellart, *professeur* at Télécom ParisTech, in the DBWeb team[1] of the *Computer science and networking* department at Télécom ParisTech. A yearly 64hr teaching mission (centered around Web, formal languages, and data management topics) is proposed in parallel with the PhD research.

References

[AAM14] Antoine Amarilli, Yael Amsterdamer, and Tova Milo. On the complexity of mining itemsets from the crowd using taxonomies. In *Proc. ICDT*, Athens, Greece, 2014. `http://arxiv.org/abs/1312.3248`.

[ABDS13] Antoine Amarilli, Lamine M. Ba, Daniel Deutch, and Pierre Senellart. Provenance for nondeterministic order-aware queries. Submitted to PODS 2014. Preprint: `http://a3nm.net/publications/amarilli2014provenance.pdf`, 2013.

[Ama13] Antoine Amarilli. Open-world query answering under number restrictions. Submitted to PODS 2014. Preprint: `http://a3nm.net/publications/amarilli2014open.pdf`, 2013.

[Ama14] Antoine Amarilli. The possibility problem for probabilistic XML (extended version). Submitted to the Alberto Mendelzon International Workshop on Foundations of Data Management, 2014. Preprint: `http://a3nm.net/publications/amarilli2014possibility.pdf`, 2014.

[AS13] Antoine Amarilli and Pierre Senellart. On the connections between relational and XML probabilistic data models. In *Proc. BNCOD*, pages 121–134, Oxford, United Kingdom, July 2013.

[BGS11] Michael Benedikt, Georg Gottlob, and Pierre Senellart. Determining relevance of accesses at runtime. In *Proc. PODS*, pages 211–222, Athens, Greece, June 2011.

[KS13] Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and complexity. In Zongmin Ma and Li Yan, editors, *Advances in Probabilistic Databases for Uncertain Information Management*, pages 39–66. Springer-Verlag, May 2013.

[SORK11] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[1] `http://dbweb.enst.fr/`