



Deliverable D4.1
First Year Report on WP 4
Dissemination level: PU
Date: 24 August 2010

1 Overview

Objectives of WP4 The broad objective is to develop a framework to deal with missing data, in particular to develop new techniques and algorithms for handling

1. missing values in XML documents through an approach based on uncertainty, and
2. missing data through automatic recovery.

The two objectives are represented by tasks T4.1 (develop a foundational framework for dealing with missing data in XML) and T4.2 (develop a foundational framework for recovering of missing metadata), respectively. Task T4.3 (integrate a prototype implementation of the new algorithms into the Software Library of T1.1 of WP1) is scheduled to start later.

Main results The key achievements of the first year are on (i) modelling uncertainty in XML, (ii) tractability of the main computational tasks, and (iii) regular expression inference. The first key achievement includes comprehensive analyses of expressiveness and succinctness of probabilistic XML models based on recursive Markov Chains, and of the interaction of incompleteness and constraints. The second achievement includes efficient algorithms for query evaluation on incomplete XML, and exact and approximate query evaluation algorithms on probabilistic data. As both achievements discuss modelling aspects of missing data as well as the developments of new tools and algorithms, they contribute to milestones I and III mentioned above. The third main achievement addresses the issue of missing data through automatic recovery. In particular, we developed new methods and tools for deterministic regular expression inference which constitutes one of the cornerstones for XML Schema (XSD) inference. As the latter will be included in the schema library, the third main achievement contributes to II and III of the project milestones.

Dissemination The results are published in [1, 6, 8, 2, 11, 12, 7, 4, 5]. The full version of [1] is under submission to the Journal of the ACM, the top journal for computer science research. Paper [3] is under submission. Publication venues include internationally leading database and

theory conferences (ACM SIGMOD, ACM PODS, IEEE ICDE, EDBT) and journals (ACM TODS, Inf Proc Letters). Talks on incomplete data management were given at several universities, as well as two symposia and two workshops. Based on contributions to probabilistic data management, we have delivered two keynote talks at international workshops [9, 10], a keynote talk at EPFL Summer Research Institute (June 2009, Lausanne), and an invited talk at Technical University of Vienna (September 2009). A 3-hour tutorial on incomplete and probabilistic data management has been given at the first FoX training week in October 2009.

Collaborations The papers on incomplete data management involved collaboration between three FoX sites: Edinburgh (David, Libkin, Reutter, and Murlak in the beginning of the year), Paris (Sirangelo), and Warsaw (Murlak for the rest of the year). It also involved an external collaborator (Barceló, University of Chile). Two of the four papers on probabilistic data management are collaborations of the Oxford site (Olteanu) and top US academic institutions Washington (Suciu) and and Cornell (Koch), thereby further increasing the awareness of this task’s contributions on the international scene. Paper [3] is a collaboration between the Oxford site (Benedikt, Olteanu) and a member of the Webdam research project funded by an ERC advanced grant (Senellart). Both papers on task T4.2 are collaborations between two FoX sites: Hasselt (Bex, Gelade, Neven, Vansummeren) and Dortmund (Martens, Schwentick).

Justification of the resources FoX funding has been used to support research positions in this project at different FoX sites and to support travel to conferences where WP4 research papers have been published and presented.

A total amount of 73 person-month (PM) has been assigned to WP4. This year we have devoted 21.05 PM to WP4 as detailed below. The fact that we used less than one-third of the PMs associated with WP4 is explained by the fact that the research positions could not be filled immediately at the start of the project.

Robert Fink was appointed by Oxford at the end of October 2009 to work on FoX as part of his Ph.D. He is working with Dan Olteanu on probabilistic data management, which is part of task T4.1.

Geert Jan Bex was appointed by Hasselt during 6 months to work as software Quality Manager for FoXLib.

In Hasselt (UHAS) Geert Jan Bex, Timos Antonopoulos and Tomasz Idziaszek were paid on FoX money and contributed a total of 11.35 PM to task T4.2.

Leonid Libkin in UEDIN also contributed 2.7 PM on WP4.

2 Description of the new results

2.1 Task 4.1

Research in this task can be classified in two broad categories, depending on whether confidences about the level of uncertainty in the data are present or not. The latter approach is called incomplete data management, whereas the former is called probabilistic data management.

Incomplete Data The main contributions to the incomplete data management are four papers: [1, 6] published in the top database theory conference, [8] from a well-known international journal, and a workshop followup [2].

The key paper is [1]: it defined a model of XML with missing information, which can be either missing data values, or missing structural information. It presented a complete classification of models of incompleteness based on which information could be missing, and studied the three main tasks that arise in such a setting:

- Consistency of incomplete descriptions: whether an incomplete description and perhaps some schema information could correspond to a XML document.
- Membership: does an incomplete description represent a complete XML document.
- Query answering: how does one compute certain answers to a query over an incomplete document.

The key goal of the paper was to draw the feasibility boundary, i.e., separate tractable cases from the intractable ones, thereby providing a list of features of incomplete XML descriptions that can be efficiently used. It did so for the three main tasks.

In [8], we looked at a special case of incompleteness when nodes come with explicit node ids. In this case, solving consistency or query answering amounts to handling some problems with related to pattern matching and pattern implication on strings. For example, query answering asks whether it is true that whenever a set of patterns is satisfied in a string in a way that patterns do not overlap, another pattern is satisfied. We provided a complete complexity analysis of these questions (in particular, we had a polynomial time algorithm for the problem just mentioned).

In [6], we extended the notion of certain answers from simple XML queries (XML-to-relations) to complex ones (XML-to-XML). We developed a general theory of certain answers, and applied it to XML data exchange (WP2), as well as settling some problems that were left open in [1].

Finally, in [2], we studied the interaction of incompleteness and schema integrity constraints in XML, providing a dichotomy for the consistency problem, fully classifying tractable and intractable cases of the problem.

Probabilistic Data The contributions to probabilistic databases are three research papers published in internationally leading conferences [11, 12, 7], and one research paper under submission [3].

In [3], we define a space of probabilistic XML models based on (restrictions of) Recursive Markov Chains and study their expressiveness, succinctness, and tractability of MSO and XPath query evaluation w/o unit-cost arithmetic. In contrast to existing probabilistic XML models, these new models (i) capture probabilistic versions of DTDs (XML schemas), (ii) can define infinite probability distributions, and (iii) can be exponentially more succinct, while still preserving MSO tractability. In addition, our tractability results subsume all known tractability results for existing probabilistic XML models.

In [7], we extend the notion of tractable queries on probabilistic databases to tractable data-query instances, where general hard queries can become tractable on restricted data instances. The query evaluation is separated into two steps: (1) the possibly large tractable data-query instance is evaluated first by efficient database techniques, and then (2) the (usually small) intractable residue is processed using inference techniques based on treewidth.

In [12], we introduce a deterministic approximation algorithm with error guarantees for confidence computation of positive relational algebra queries in (arbitrarily correlated) probabilistic databases. An important property of this algorithm is that, without explicit knowledge of the query, it provably finishes in polynomial time for known tractable queries. This algorithm is based on incremental compilation of query lineage into so-called decomposition trees that support linear-time probability computation. The compilation is incremental and we show experimentally that it can achieve a given approximation within a few steps. We implemented the algorithm in the SPROUT query engine, which is publicly available at <http://www.comlab.ox.ac.uk/projects/SPROUT/>.

The paper [11] is the first to discuss tractability of conjunctive queries with inequalities in probabilistic databases. Conjunctive queries with inequalities subsume XML queries such as those defined by the navigational XPath language. It presents a class of hard queries, and gives a scalable algorithm for tractable inequality queries that is based on ordered binary decision diagrams. This algorithm is implemented in the query engine SPROUT and passed the SIGMOD'09 repeatability and workability evaluation (RWE); out of 63 accepted papers, 19 participated in RWE, and 10 passed RWE.

In the years to come, we would like to investigate: (a) the data com-

plexity of conjunctive queries with inequalities, which subsume standard navigational XML queries, is not completely charted (are they either tractable or $\#P$ -hard, or is there something in between?), (b) exact and approximate query evaluation algorithms for full relational algebra and XPath (including difference) are still to be found, (c) randomised approximation schemes for query evaluation in richer models, such as those based on recursive Markov Chains, are still to be investigated, (d) query answering in expressive languages over incomplete documents is still poorly understood, and (e) the role of constraints deserve a much more thorough investigation.

2.2 Task 4.2

The main focus of this task is on XSD recovery. Basically the latter problem reduces to two interrelated sub problems (1) learning of regular expressions constituting the content models of schemata: and (2) learning of types. With respect to the first problem, we devised new algorithm to translate automata to regular expressions. The presence of such methods allows to use existing automata inference algorithms to learn an automata from a given sample of strings, which can then be subsequently rewritten into a regular expression. Existing methods are based on state elimination and produce an equivalent regular expression which can be exponentially larger. To obtain an equivalent regular expression whose size is small on average, we developed a new method based on a structural decomposition of the automaton. This algorithm is currently implemented but needs to be tested extensively. When the algorithm is successful, a paper containing the newly derived results will be written. With respect to the second problem, i.e., the learning of types, we are adopting a hidden markov model approach to infer probabilistic XSDs. However, to transform such a probabilistic XSD to a standard deterministic one, we need a methodology to compare them. That is, given two XSDs we need to compute the similarity between them. The latter turned out to be a computational bottleneck. We have been and still are investigating techniques from analytical combinatorics to circumvent the bottleneck. A paper on this topic is in preparation.

In the years to come, we will continue working on these topics. We did not start yet to investigate inference of XML transformations. This research will be started in the upcoming year. Task 4.3 consists of adding the new learning algorithms to FoXLib. As the implementation of the core of FoXLib is scheduled for the coming year, we will start adding the new algorithms when the core functionality of the library is available.

References

- [1] P. Barceló, L. Libkin, A. Poggi, and C. Sirangelo. XML with incomplete information: models, properties, and query answering. In *Proc. of ACM Symp. on Principles of Database Systems (PODS), Lausanne*, pages 237–246, 2009.
- [2] P. Barceló, L. Libkin, and J. Reutter. On incomplete XML documents with integrity constraints. In *Proc. of the 4th Alberto Mendelzon Workshop on Foundations of Data Management*, 2010.
- [3] M. Benedikt, E. Kharlamov, D. Olteanu, and P. Senellart. Probabilistic XML via markov chains. submitted to VLDB 2010.
- [4] G. J. Bex, W. Gelade, W. Martens, and F. Neven. Simplifying XML schema: Effortless handling of nondeterministic regular expressions. In *ACM SIGMOD*, pages 731–744, 2009.
- [5] G. J. Bex, F. Neven, T. Schwentick, and S. Vansummeren. Inference of concise regular expressions and DTDs. *ACM Trans. Database Syst.*, 35(2), 2010.
- [6] C. David, L. Libkin, and F. Murlak. Certain answers for XML queries. In *Proc. of ACM Symp. on Principles of Database Systems (PODS), Lausanne*, 2010.
- [7] A. Jha, D. Olteanu, and D. Suciu. Bridging the gap between intensional and extensional query evaluation in probabilistic databases. In *Proc. of Int. Conf. on Extending Data Base Technology (EDBT), Lausanne*, 2010.
- [8] L. Libkin and C. Sirangelo. Disjoint pattern matching and implication in strings. *Information Processing Letters*, 110(4):143–147, 2010.
- [9] D. Olteanu. Keynote talk: A toolbox of query evaluation techniques for probabilistic databases. In *Workshop on Logic in Databases (LID), Copenhagen*, 2009.
- [10] D. Olteanu. Keynote talk: A toolbox of query evaluation techniques for probabilistic databases. In *IEEE ICDE workshop on Managing and Mining of Uncertain Data (MOUND), Long Beach*, 2010.
- [11] D. Olteanu and J. Huang. Secondary-storage confidence computation for conjunctive queries with inequalities. In *ACM Special Interest Group on Management of Data (SIGMOD), Providence*, 2009.
- [12] D. Olteanu, J. Huang, and C. Koch. Approximate confidence computation in probabilistic databases. In *Proc. of IEEE Int. Conf. on Data Engineering (ICDE), Long Beach*, 2010.