IN PARTNERSHIP WITH:
CNRS

Ecole normale supérieure de Cachan

Activity Report 2012

Project-Team DAHU

Database and verification

IN COLLABORATION WITH: Laboratoire specification et vérification (LSV)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Knowledge and Data Representation and Management

Table of contents

Project-Team DAHU

Keywords: Data Management, Databases, Web, Verification, Distributed System

Dahu is a common project with CNRS and ENS de Cachan located in the LSV. The team was created on January the 1st, 2008.

Creation of the Project-Team: January 01, 2009 .

1. Members

Research Scientists

Serge Abiteboul [Senior researcher, Inria, HdR]
Luc Segoufin [Team leader, Senior Researcher, Inria, HdR]

Faculty Member

Cristina Sirangelo [Associate Professor, ENS Cachan]

PhD Students

Émilien Antoine [WebDaM]
Nadime Francis [ENS-Cachan then ASN]
Wojciech Kazana [WebDaM]

Post-Doctoral Fellow

M Praveen [ERCIM]

2. Overall Objectives

2.1. Introduction

For more information see http://www.lsv.ens-cachan.fr/axes/DAHU/dahu.php.

The need to access and exchange data on the Web has led to database management systems (DBMS) that are increasingly distributed and autonomous. Data extraction and querying on the Web is harder than in classical DBMS, because such data is heterogeneous, redundant, inconsistent and subject to frequent modifications. DBMS thus need to be able to detect errors, to analyze them and to correct them. Moreover, increasingly complex Web applications and services rely on DBMS, and their reliability is crucial. This creates a need for tools for specifying DBMS in a high-level manner that is easier to understand, while also facilitating verification of critical properties.

The study of such specification and verification techniques is the main goal of Dahu.

2.2. Highlights of the Year

Serge Abiteboul has been professor at College de France till September 2012. He organized a course on Web data management.

3. Scientific Foundations

3.1. Scientific Foundations

Dahu has strong connections with the Leo project-team in Saclay.

Dahu aims at developing mechanisms for high-level specifications of systems built around DBMS, that are easy to understand while also facilitating verification of critical properties. This requires developing tools that are suitable for reasoning about systems that manipulate data. Some tools for specifying and reasoning about data have already been studied independently by the database community and by the verification community, with various motivations. However, this work is still in its infancy and needs to be further developed and unified.

Most current proposals for reasoning about DBMS over XML documents are based on tree automata, taking advantage of the tree structure of XML documents. For this reason, the Dahu team is studying a variety of tree automata. This ranges from restrictions of "classical" tree automata in order to understand their expressive power, to extensions of tree automata in order to understand how to incorporate the manipulation of data.

Moreover, Dahu is also interested in logical frameworks that explicitly refer to data. Such logical frameworks can be used as high level declarative languages for specifying integrity constraints, format change during data exchange, web service functionalities and so on. Moreover, the same logical frameworks can be used to express the critical properties we wish to verify.

In order to achieve its goals, Dahu brings together world-class expertise in both databases and verification.

4. Application Domains

4.1. Application Domains

Databases are pervasive across many application fields. Indeed, most human activities today require some form of data management. In particular, all applications involving the processing of large amounts of data require the use of a database. Increasingly complex Web applications and services also rely on DBMS, and their correctness and robustness is crucial.

We believe that the automated solutions that Dahu aims to develop for verifying such systems will be useful in this context.

5. New Results

5.1. Distributed data management

Participants: Serge Abiteboul, Émilien Antoine, Cristina Sirangelo, Nadime Francis, Luc Segoufin.

Distributed knowledge base. We are developing the system Webdamlog [16], [13], [14] to address the challenges faced by everyday Web users, who interact with inherently heterogeneous and distributed information. Managing such data is currently beyond the skills of casual users. In Webdamlog, we see the Web as a knowledge base consisting of distributed logical facts and rules. The objective is to enable automated reasoning over this knowledge base, ultimately improving the quality of service and of data. The system supports the Webdamlog language, a Datalog style language with rule delegation.

Deduction in uncertain worlds. Motivated by reasoning in distributed environments in which disagreements arise between different actors, we study in [17] deduction (captured by datalog programs) in the presence of inconsistencies (induced by functional dependency (FD) violations). We adopt an operational semantics for datalog with FDs based on inferring facts one at a time, while never violating the FDs. This yields a set of possible worlds that we capture by c-tables of possibly exponential size. We propose to use probabilities to measure this nondeterminism and define a probabilistic semantics that can be captured by probabilistic conditional tables. Not surprisingly, we show that computing the probability of a query answer in our setting is expensive, which leads us to introduce a sampling algorithm to estimate answer probabilities. We then turn our attention to the problem of explaining why a particular answer holds. This leads us to consider two novel notions: the most influential

extensional facts, and the most likely proofs for an answer. We study algorithms for ranking facts and proofs based on their contribution to the derivation of an answer. Finally, we consider how our framework can be adapted to a distributed setting, and in particular, how sampling can be performed in a distributed manner.

Access rights in a distributed setting. We started considering access right issues in Webdamlog. This is related to specifying access right on views in standard databases. There is also the issues of controlling rules that are run locally but were specified by other peers.

Incomplete information in Web data. Incomplete information often arises from the integration of different Web data sources, as well as from the exchange of data between communicating Web applications. The semantics of incompleteness (i.e. which possible complete databases are represented by an incomplete one) depends on the context and the particular scenario where incompleteness raises from. We have studied how to deal with the presence of incomplete information under different possible semantics. We have in particular studied in which condition it is possible to query incomplete data "naively", i.e. as if it were complete. We have exhibited "natural" fragments of first order logic for which naive evaluation is possible, under different semantics.

Graph data management. Graph structured data can be found in new emerging applications such as RDF and linked data, or social networks. The peculiarity of queries over graphs is that they are interested in both data carried by the graph and in the graph topology; they are often based on reachability patterns. In a distributed setting it is very common to be able to query only a partial description or a "view" of the graph. We studied the problem of answering queries using only the information provided by the views. The presence of a form of recursion in views and queries presents new challenges. We found restricted classes of graph views and queries that allow efficient query answering over views.

5.2. Tree automata theory

Participants: Luc Segoufin, Serge Abiteboul, M Praveen.

Tree automata We studied the expressive power of a subclass of regular tree languages. We gave a decidable characterization of those languages that are "piecewise testable", i.e. definable using boolean combination of existential first-order formulas [12].

Automata with counters. We studied extending techniques used in standard Petri nets to other models. We extended the Rackoff technique to decide coverability and boundedness problems for Strongly Increasing Affine nets, a subclass of Affine nets [20].

Languages on trees. We studied in [18] highly expressive query languages for unordered data trees, using as formal vehicles Active XML and extensions of languages in the while family. All languages may be seen as adding some form of control on top of a set of basic pattern queries. The results highlight the impact and interplay of different factors: the expressive power of basic queries, the embedding of computation into data (as in Active XML), and the use of deterministic vs. nondeterministic control. All languages are Turing complete, but not necessarily query complete in the sense of Chandra and Harel. Indeed, we show that some combinations of features yield serious limitations, analogous to FOk definability in the relational context. On the other hand, the limitations come with benefits such as the existence of powerful normal forms. Other languages are "almost" complete, but fall short because of subtle limitations reminiscent of the copy elimination problem in object databases.

Probabilistic XML. In [15], we study the problem of, given a corpus of XML documents and its schema, finding an optimal (generative) probabilistic model, where optimality here means maximizing the like- lihood of the particular corpus to be generated. Focusing first on the structure of documents, we present an efficient algorithm for finding the best generative probabilistic model, in the absence of constraints. We further study the problem in the presence of integrity constraints, namely key, inclusion, and domain constraints. We study in this case two different kinds of generators. First, we consider a continuation-test generator that performs, while generating documents, tests of schema

satisfiability ; these tests prevent from generating a document violating the constraints but, as we will see, they are computationally expensive. We also study a restart generator that may generate an invalid document and, when this is the case, restarts and tries again. Finally, we consider the injection of data values into the structure, to obtain a full XML document. We study different approaches for generating these values.

Infinite alphabet. We studied the complexity of satisfiability of linear temporal logics extended to reason about repetitions of values from an infinite data domain. We refined an existing result that reduced this problem to Petri net reachability, and showed that it can be reduced to the coverability problem. Using this refinement, we gave the precise complexity of the satisfiability problem. We also characterized the complexity of satisfiability for many fragments and extensions of the logic.

6. Partnerships and Cooperations

6.1. European Initiatives

6.1.1. FP7 Projects

6.1.1.1. FOX

Title: FOX

Type: COOPERATION (ICT)

Defi: FET Open

Instrument: Specific Targeted Research Project (STREP)

Duration: May 2009 - September 2012

Coordinator: Inria (France)

Others partners: Thomas Schwentick at the university of Dortmund, Mikołaj Bojańczyk at the university of Warsaw, Leonid Libkin at the university of Edinburgh, Georg Gottlob at the university of Oxford, Frank Neven at the university of Hasselt and Maarten Marx at the university of Amsterdam.

See also: http://fox7.eu

Abstract: The objective of FoX is to study the fundamental issues necessary in order to make the data management over the internet more efficient and more reliable.

6.1.2. ERC grant Webdam

6.1.2.1. Webdam

Title: WebDam

Type: IDEAS

Instrument: ERC Advanced Grant (Advanced)

Duration: December 2008 - November 2013

Coordinator: Serge Abiteboul, Inria (France)

Others partners: Pierre Senellart, Telecom Paristech.

See also: http://webdam.inria.fr

Abstract: The goal is to develop a formal model for Web data management. This model will open new horizons for the development of the Web in a well-principled way, enhancing its functionality, performance, and reliability. Specifically, the goal is to develop a universally accepted formal framework for describing complex and flexible interacting Web applications featuring notably data exchange, sharing, integration, querying and updating. We also propose to develop formal foundations that will enable peers to concurrently reason about global data management activities, cooperate in solving specific tasks and support services with desired quality of service.

6.2. International Initiatives

6.2.1. Inria International Partners

- Victor Vianu, UC San Diego, USA.

6.3. International Research Visitors

6.3.1. Visits of International Scientists

Victor Vianu (from June 2012 until September 2012)

Subject: WebDaM

Institution: UC San Diego (USA)

Gerome Miklau (from September 2012 to June 2012)

Subject: WebDaM

Institution: Univesity of Massachusetts at Amherst (USA)

6.3.2. Internships

- David Montoya, Webdam, 04/2012 to 09/2012
- Jules Testard, Webdam, 09/2012 to 12/2012

7. Dissemination

7.1. Scientific Animation

Organization of workshops and conferences.

- Luc Segoufin organized an international workshop on finite model theory in Les Houches in May 14-18, 2012: http://www.lsv.ens-cachan.fr/Events/fmt2012/.
- Serge Abiteboul co-organized with Tova Milo (Tel Aviv) the WebDam-MoDaS Workshop on Web data management and Crowdsourcing, Eilat, Israel 2012
- Serge Abiteboul co-organized with P. Senellart (Telecom Paris) the Webdam "Data in the Wild" Workshop, Paris 2012

Program Committees.

- Serge Abiteboul: International Conference on Database Theory (ICDT'12)
- Serge Abiteboul: World Wide Web (WWW'2012)
- Luc Segoufin: Principle of Database systems (PODS'12).
- Cristina Sirangelo: International Conference on Database Theory (ICDT'13)
- Cristina Sirangelo: ICDT'13 Test of Time Award
- Cristina Sirangelo: 6th Alberto Mendelzon Workshop

Responsibilities.

- Luc Segoufin is since 2009 the coordinator of the european project FoX. Since 2010 he is a member of the steering committee of the Intl. Conf. on Database Theory (ICDT). Since 2010 he is part of the "bureau du commité des projets" à l'Inria Saclay. Since 2011 he is part of the scientific board of Inria. Since 2010 he is responsible of the groupe de travail "Complexité et Modèles Finis" du GDR "Mathématique et Informatique" (http://www.gdr-im.fr/).

– Serge Abiteboul is the principal investigator of the European Research Council Grant Webdam on Web Data Management. He is a member of the French Academy of Sciences and of the Academia Europea. He is chairman of the Scientific Council of Société d'Informatique de France, elected in 2012.

Larger audience. School "Imagine the Future in ICT", organized by ICDT lab. Two courses given by Serge Abiteboul: (i) Data sciences; (ii) Web search engines.

Nomination. Serge Abiteboul has been professor at College de France [21] till September 2012. He organized a 10 hours course on Web data management. He also organized a seminar on the topic with for guests: Moshe Vardi, Anastasia Ailamaki, François Bancilhon, Julien Masanès, Victor Vianu, Tova Milo, Georg Gottlob, Gerhard Weikum, Marie-Christine Rousset, Pierre Senellart.

7.2. Teaching - Supervision - Juries

7.2.1. Teaching

Le module "Teaching" présente vos activités d'enseignement et d'encadrement à présenter comme suit :

Licence : Cristina Sirangelo, Bases de données, 30 heures équivalent TD, L3, École Normale Supérieure de Cachan, France

Doctorat : Cristina Sirangelo, Bases de données et sites Web dynamiques, 18 heures équivalent TD, École Normale Supérieure de Cachan, France

Doctorat : Cristina Sirangelo, Création de sites Web, 18 heures équivalent TD, École Normale Supérieure de Cachan, France

Licence : Serge Abiteboul, Base de données , ENS Cachan and ENS Paris

Master : Serge Abiteboul, Web data management, MPRI Paris

Licence : Émilien Antoine, Algorithmique et complexité, 32h, L3, Université de Paris-Sud, France

7.2.2. Supervision

PhD & HdR :

PhD in Progress: Wojciech Kazana, enumeration of queries, 01/03/2010, Luc Segoufin

PhD in Progress: Nadime Francis, graph databases, 01/09/2011, Cristina Sirangelo and Luc Segoufin

PhD in progress : Émilien Antoine, Data management in social network, 01/10/2010, Serge Abiteboul

7.2.3. Juries

- Luc Segoufin was on the HDR jury of Florent Madeleine at the university of Clermont-Ferrand.
- Luc Segoufin was on the HDR jury of Manuel Bordiski at the university of Paris 7.
- Cristina Sirangelo was on the jury for a MCF position at Ecole Normale Superieure de Cachan

7.3. Popularization

S.Abiteboul's radio talk shows: Science Publique (France Culture), Place de La Toile (France Culture), Autour de la question (RFI)

S.Abiteboul's interviews in the press:

- Construisons un Web des savoirs, Le Monde
- Le Web redéfinit sans cesse les échanges d'information, Le Figaro 2012
- L'enseignement de l'informatique en classes prépas, 01.Net (avec Colin de La Higuerra)
- Le Big Data est avant tout un effet de mode, O1Net
- Sur les liens entre labos publics et universitaires, 01.Net 2012.
- L'informatique est une science bien trop sérieuse pour être laissée aux informaticiens, Le monde.fr (avec Colin de la Higuera et Gilles Dowek)
- L'important sur Internet, c'est de trouver la bonne information, Lepoint.fr

8. Bibliography

Major publications by the team in recent years

[1] S. ABITEBOUL, I. MANOLESCU, P. RIGAUX, M.-C. ROUSSET, P. SENELLART. *Web Data Management*, Cambridge University Press, 2012, 456, http://hal.inria.fr/hal-00677720.

[2] S. ABITEBOUL, L. SEGOUFIN, V. VIANU. *Static Analysis of Active XML Systems*, in "ACM Transactions on Database Systems", 2009, vol. 34, n° 4.

[3] P. BARCELÓ, L. LIBKIN, A. POGGI, C. SIRANGELO. *XML with incomplete information*, in "J. ACM", 2010, vol. 58, n° 1.

[4] M. BOJANCZYK, L. SEGOUFIN, H. STRAUBING. *Piecewise testable tree languages*, in "Logical Methods in Computer Science (LMCS)", 2012, vol. 8, n° 3.

[5] M. BOJAŃCZYK, C. DAVID, A. MUSCHOLL, T. SCHWENTICK, L. SEGOUFIN. *Two-variable logic on words with data*, in "ACM Trans. on Computational Logic (ToCL)", 2011, vol. 12, n° 4.

[6] M. BOJAŃCZYK, A. MUSCHOLL, TH. SCHWENTICK, L. SEGOUFIN. *Two-variable logic on data trees and applications to XML reasoning*, in "Journal of the ACM", 2009, vol. 56, n° 3.

[7] BALDER TEN. CATE, L. SEGOUFIN. *Transitive Closure Logic, Nested Tree Walking Automata, and XPath*, in "Journal of the ACM", 2010, vol. 57, n° 3.

[8] B. CAUTIS, S. ABITEBOUL, T. MILO. *Reasoning about XML update constraints*, in "Journal of Computer and System Sciences", 2009, vol. 75, n° 6, p. 336-358.

[9] L. LIBKIN, C. SIRANGELO. *Reasoning about XML with temporal logics and automata*, in "Journal of Applied Logic", 2010, vol. 8, n° 2, p. 210-232, http://www.lsv.ens-cachan.fr/Publis/PAPERS/PDF/LS-jal10.pdf.

[10] L. LIBKIN, C. SIRANGELO. *Data exchange and schema mappings in open and closed worlds*, in "Journal of Computer System Sciences (JCSS)", 2011.

Publications of the year

Articles in International Peer-Reviewed Journals

[11] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Comparing workflow specification languages: A matter of views*, in "ACM Trans. on Database systems (ToDS)", 2012, vol. 37, n° 2, 10.

[12] M. BOJANCZYK, L. SEGOUFIN, H. STRAUBING. *Piecewise testable tree languages*, in "Logical Methods in Computer Science (LMCS)", 2012, vol. 8, n° 3.

International Conferences with Proceedings

[13] S. ABITEBOUL. *Sharing Distributed Knowledge on the Web (Invited Talk)*, in "Proc. of Computer Science Logic (CSL)", 2012, p. 6-8.

[14] S. ABITEBOUL. *Viewing the Web as a Distributed Knowledge Base*, in "Description Logics", 2012.

[15] S. ABITEBOUL, Y. AMSTERDAMER, D. DEUTCH, T. MILO, P. SENELLART. *Finding optimal probabilistic generators for XML collections*, in "Proc. of Intl. Conf. on Database Theory (ICDT)", 2012, p. 127-139.

[16] S. ABITEBOUL, E. ANTOINE, J. STOYANOVICH. *Viewing the Web as a Distributed Knowledge Base*, in "Proc. of Intl. Conf. on Database Engineering (ICDE)", 2012, p. 1-4.

[17] S. ABITEBOUL, M. BIENVENU, D. DEUTCH. *Deduction in the Presence of Distribution and Contradictions*, in "WebDB", 2012, p. 31-36.

[18] S. ABITEBOUL, P. BOURHIS, V. VIANU. *Highly expressive query languages for unordered data trees*, in "Proc. of Intl. Conf. on Database Theory (ICDT)", 2012, p. 46-60.

[19] S. ABITEBOUL, P. SENELLART, V. VIANU. *The ERC webdam on foundations of web data management*, in "WWW (Companion Volume)", 2012, p. 211-214.

[20] R. BONNET, A. FINKEL, M. PRAVEEN. *Extending the Rackoff technique to Affine nets*, in "Proceedings of the IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FST-TCS)", Schloss Dagstuhl –Leibniz Center for Informatics, 2012.

Scientific Books (or Scientific Book chapters)

[21] S. ABITEBOUL. *Sciences des données: de la Logique du premier ordre à la Toile*, Fayard, 2012, http://abiteboul.com/College/lecon.htm.

[22] S. ABITEBOUL, I. MANOLESCU, P. RIGAUX, M.-C. ROUSSET, P. SENELLART. *Web Data Management*, Cambridge University Press, 2012, 456, http://hal.inria.fr/hal-00677720.