ERC Webdam

Acti it Report 2012

1 Executive summary

2 The general goal

3 Participants

Members

- Serge Abiteboul, manager of the Webdam ERC project
- Gerome Miklau, U. Mass. Amherst [since 09/2012]
- Victor Vianu, professor at UC San Diego [07/2010-07/2011]

PhD Students

- Emilien Antoine, Paris Sud University
- Wojciech Kazana, ENS Cachan
- Marilena Oita, Telecom ParisTech [till 01/11/2012] NOT DAHU

Others

- Jules Testard, undergraduate McGill, Canada [09/2012-12/2012]
- David Montoya, master MPRI [04/2012-09/2012]

Collaborators

- Meghyn Bienvenu, researcher at Orsay University NOT DAHU
- Luc Segoufin, manager of Dahu team, INRIA Saclay
- Pierre Senellart, associate professor at Telecom ParisTech NOT DAHU
- Julia Stoyanovich, Drexel U., USA NOT DAHU

Visitors

- [**Émilien** visitors are people that stays at least one month ? YES]

4 Research

List of main results obtained in 2012.

Distributed knowledge base. (Serge, Émilien, Julia) We are developing the system Webdamlog [5, 1, 2] 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.

Probabilistic XML. (Serge) In [3], 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 likelihood 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.

Based on these results, we developed ALEX, an Auto-completion Learning Editor for XML. The editor assists the users by providing intelligent auto-completion suggestions.

Languages on trees. (Serge, Victor) We studied in [8] 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.

Deduction in uncertain worlds. (Serge) Motivated by reasoning in distributed environments in which disagreements arise between different actors, we study in [6] 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. (Gerome, Serge, Émilien, Julia) 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.

5 Dissemination

5.1 PhD Thesis

- PhD : Marilena Oita, Deriving Semantic Objects from the Structured Web, Télécom ParisTech, 14/10/2012, Pierre Senellart

5.2 Program committees

- Serge Abiteboul:
  - International Conference on Database theory, Berlin, 2012

  - World Wide Web, Lyon, 2012

- Pierre Senellart
- Victor Vianu

5.3 Organization of workshops and conferences

Serge Abiteboul co-organized with Tova Milo (Tel Aviv) the WebDam-MoDaS Workshop on Web data management and Crowdsourcing, Eilat, Israel 2012

Emilien Antoine co-organized the brainstorming session at that workshop

Serge Abiteboul co-organized with P. Senellart (Telecom Paris) the Webdam "Data in the Wild" Workshop, Paris 2012

5.4 Education

Abiteboul published "Sciences des données", Fayard 2012. The book is available on the Web at http://lecons-cdf.revues.org/506 and in English translation by Liz Libbrecht at http://lecons-cdf.revues.org/558.

Serge Abiteboul's teaching:

- Serge Abiteboul has been professor at College de France 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.
- School "Imagine the Future in ICT", organized by ICDT lab. Two courses: (i) Data sciences; (ii) Web search engines.
- Relational databases, undergratuate course, ENS Cachan and ENS Paris.
- Web data management, graduate course, MPRI Paris.

  [**Émilien** should I add my teaching here]

5.5 Invited Presentations and tutorials

S. Abiteboul:

- Sharing Distributed Knowledge on the Web, Conference on Computer Science Logic, Fontainebleau 2012

- Viewing the Web as a Distributed Knowledge Base, International Workshop on Description Logic, Roma 2012
- Web data management, INTIMATE workshop on "Big Data in Digital Life", Paris 2012
- Life in Academia and How to Choose a Thesis Topic, First AVSE Doctoral Workshop, Cachan 2012
- Science of data: from first order logic to the Web, Colloque « Translittératies : enjeux de citoyenneté et de créativité » ENS-Cachan et Université Sorbonne nouvelle, Cachan 2012
- Collective question answering, MSR-INRIA Workshop, Cambridge 2012
- Données, Information, connaissances : challenges, Data Excellence Paris Conference, Paris 2012
- La gestion de données à l'heure de la Toile, Data Tuesday, (avec Fernando Velez), Issy-les-Moulineaux 2012
- Quelle cuisine pour les données du Web ? Let's imagine the Future workshop, Rennes 2012
- Overview of Webdam, WebDam-MoDaS Workshop, Eilat, Israel 2012
- How can humans and systems collaborate in a social network to answer queries: issues and challenges, panel with P. Buneman, M. Franklin, H.V. Jagadish
- Viewing the Web as a Distributed Knowledge Base, French-Israeli Workshop on Foundations of Computer Science, Paris 2012
- Viewing the Web as a Distributed Knowledge Base, EPFL, Lausanne 2012

É. Antoine:

- Access control in WebdamLog system, WebDam-MoDaS Workshop, Eilat, Israel 2012
- Managing Distributed Knowledge on the Web, BDA, Clermont-Ferrant 2012.

5.6 Journal editing

- Victor Vianu: Editor-in-chief of JACM, Area editor for ACM Trans. on Computational Logic (logical aspects of databases), Editor of the Database Theory Column of SIGACT News
- Pierre Senellart is Information Director of the Journal of the ACM. NOT DAHU

5.7 Responsibilities

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.

5.8 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

References

[1] Serge Abiteboul. Sharing distributed knowledge on the web (invited talk). In *CSL*, pages 6–8, 2012.

[2] Serge Abiteboul. Viewing the web as a distributed knowledge base. In *Description Logics*, 2012.

[3] Serge Abiteboul, Yael Amsterdamer, Daniel Deutch, Tova Milo, and Pierre Senellart. Finding optimal probabilistic generators for xml collections. In *ICDT*, pages 127–139, 2012.

[4] Serge Abiteboul, Yael Amsterdamer, Tova Milo, and Pierre Senellart. Auto-completion learning for xml. In *SIGMOD'12*, 2012.

[5] Serge Abiteboul, Emilien Antoine, and Julia Stoyanovich. Viewing the web as a distributed knowledge base. In *ICDE*, pages 1–4, 2012.

[6] Serge Abiteboul, Meghyn Bienvenu, and Daniel Deutch. Deduction in the presence of distribution and contradictions. In *WebDB*, pages 31–36, 2012.

[7] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Comparing workflow specification languages: A matter of views. *ACM Trans. Database Syst.*, 37(2):10, 2012.

[8] Serge Abiteboul, Pierre Bourhis, and Victor Vianu. Highly expressive query languages for unordered data trees. In *ICDT*, pages 46–60, 2012.

[9] Serge Abiteboul, Pierre Senellart, and Victor Vianu. The erc webdam on foundations of web data management. In *WWW (Companion Volume)*, pages 211–214, 2012.