The Story of Webdamlog

Serge Abiteboul

The Story of Webdamlog *

Serge Abiteboul

INRIA Saclay & ENS Cachan

Abstract. We summarize in this paper works about the management of data in a distributed manner based on Webdamlog, a datalog-extension. We point to relevant articles on these works. More references may be found there.

1 The Webdamlog approach

Information of interest may be found on the Web in a variety of forms, in many systems, and with different access protocols. Today, the control and management of the diversity of data and tasks in this setting are beyond the skills of casual users [1]. Facing similar issues, companies see the cost of managing and integrating information skyrocketing. We are concerned with the management of Web data in place in a distributed manner, with a possibly large number of autonomous, heterogeneous systems collaborating to support certain tasks. We summarize in this paper works in this setting around Webdamlog and point to the relevant articles on it.

The thesis is that managing the richness and diversity of data residing on the Web can be tamed using a holistic approach based on a distributed knowledge base. Our approach is to represent all Web information as logical facts, and Web data management tasks as logical rules. A variety of complex data management tasks that currently require intense work and deep expertise may then greatly benefit from the automatic reasoning provided by inference engines, operating over the distributed Web knowledge base: for instance, information access, access control, knowledge acquisition and dissemination.

We propose to express the peers logic in Webdamlog, a datalog-style rule-based language. In Webdamlog, peers exchange facts (for information) and rules (in place of code). The use of declarative rules provides the following advantages. Peers may perform automatic reasoning using the available knowledge. Because the model is formally defined, it becomes possible to prove (or disprove) desirable properties. Because the model is based on a datalog-style language, query processing can benefit from optimization techniques. Because the model represents provenance and time, the quality of data can be better controlled. Because

* We thank all the researchers who participated in the Webdamlog project and in particular, Meghyn Bienvenu, Pierre Bourhis, Daniel Deutch, Alban Galland, Gerome Miklau, Vera Zaychik Moffitt, Marie-Christine Rousset, Julia Stoyanovich, Jules Testard, and Victor Vianu.

the model is general, a wide variety of scenarios and protocoles may be captured, which is a requirement for todays Web.

This work was realized in the context of the European Research Council grant Webdam [13, 6]. The system is available in opensource at [8]. The work on Webdamlog was inspired by previous works on ActiveXML [3] at INRIA, as well as Bud [12, 7] at Berkeley University. The system has been demonstrated in [2]. An extensive experimental evaluation of the implementation (showing notably that the computational cost of access control is modest) is presented in [11].

In the remaining of this paper, we briefly mention three main contributions: (i) The Webdamlog language that facilitates the exchange of data and rules between distributed peers; (ii) A collaborative access control mechanism for Webdamlog that enables controlling the dissemination of data in a network; and (iii) A probabilistic semantics for datalog with functional dependencies that can serve as the basis for managing uncertain, noisy, possibly contradicting data.

2 Three main contributions

Webdamlog. There is a new trend to use datalog-style rule-based languages to specify modern distributed applications, notably on the Web [9, 10]. The Webdamlog language was first formally described in [4]. It is a version of distributed datalog that allows specifying distributed applications where peers exchange messages (i.e. logical facts) as well as rules (i.e., the analog of code). An example of rule is as follows:

```
[at alice] album@alice($photoId,$photo,$f) :- friend@alice($f),
           album@$f($photoId,$photo,$source), tags@f($photoId,"Alice")
```

Ignore the details of the syntax. With this rule, Alice deploys, at each peer corresponding to one of her friends, a rule that sends her all photos this friend owns that is tagged by her name. The main originality of the language is the use of *delegation* that allows delegating rules to other peers. Distributed computating is realized by delegating some rules to perform some tasks to other peers. Knowledge acquisition, i.e., the Webdamlog analog to "downloading apps", is also performed using rule delagations. The main contribution of [4] is the presentation of the language. A study of the impact on expressiveness of "delegations" is also provided.

Access control. Users wish to share data using these systems, but avoiding the risks of unintended disclosures or unauthorized access by applications has become a major challenge. An important issue for users in a distributed setting is thus the control of the access to their data by others. In [11], we introduce a collaborative access control mechanism for Webdamlog. Using this model, users can specify access control policies providing flexible tuple-level control derived using provenance information.

Inconsistency and imprecision. In [5], we study deduction in the presence of inconsistencies and probabilites for datalog programs. (The results can be extended to Webdamlog in a straightward manner). Inconsistencies are captured

through violations of functional dependencies (FDs). We propose nondeterministic semantics for datalog with FDs. We introduce a PTIME (in the size of the extensional data) algorithm, that given a datalog program, a set of FDs and an input instance, produces a c-table representation of the set of possible resulting worlds.

We then propose to quantify nondeterminism with probabilities, by means of a probabilistic semantics. We consider the problem of capturing possible worlds along with their probabilities via probabilistic c-tables. We then study classical computational problems in this novel context. We consider the problems of computing the probabilities of answers, of identifying most likely supports for answers, and of determining the extensional facts that are most influential for deriving a particular fact. We show that the interplay of recursion and FDs leads to novel technical challenges in the context of these problems.

References

1. S. Abiteboul, B. André, and D. Kaplan. Managing your digital life. *Commun. ACM*, 58(5):32–35, 2015.
2. S. Abiteboul, E. Antoine, G. Miklau, J. Stoyanovich, and J. Testard. [Demo] rule-based application development using WebdamLog. In *SIGMOD*, 2013.
3. S. Abiteboul, O. Benjelloun, and T. Milo. The active XML project: an overview. *VLDB J.*, 17(5):1019–1040, 2008.
4. S. Abiteboul, M. Bienvenu, A. Galland, and E. Antoine. A rule-based language for Web data management. In *PODS*, 2011.
5. S. Abiteboul, D. Deutch, and V. Vianu. Deduction with contradictions in datalog. In *International Conference on Database Theory*, 2014.
6. S. Abiteboul, P. Senellart, and V. Vianu. The ERC webdam on foundations of web data management. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*, pages 211–214, 2012.
7. P. Alvaro, N. Conway, J. Hellerstein, and W. R. Marczak. Consistency analysis in bloom: a calm and collected approach. In *CIDR*, pages 249–260, 2011.
8. E. Antoine. The webdamlog system on github. https://github.com/Emilien-Antoine/webdamlog-engine, 2013.
9. J. M. Hellerstein. Datalog redux: experience and conjecture. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–2. ACM, 2010.
10. S. S. Huang, T. J. Green, and B. T. Loo. Datalog and emerging applications: an interactive tutorial. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1213–1216. ACM, 2011.
11. V. Z. Moffitt, J. Stoyanovich, S. Abiteboul, and G. Miklau. Collaborative access control in webdamlog. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 197–211, 2015.
12. B. O. O. M. project. Bloom programming language. http://www.bloom-lang.net/.
13. The Webdam ERC Project. http://webdam.inria.fr/.