The ERC Webdam on Foundations of Web Data Management*

Serge Abiteboul
Collège de France
INRIA Saclay & ENS Cachan
fname.lname@inria.fr

Pierre Senellart
Institut Télécom; Télécom ParisTech
CNRS LTCI
fname.lname@telecom-paristech.fr

Victor Vianu
U.C. San Diego
lname@cs.ucsd.edu

ABSTRACT

The Webdam ERC grant is a five-year project that started in December 2008. The goal is to develop a formal model for Web data management that would 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. Although the proposal addresses fundamental issues, its goal is to serve as the basis for future software development for Web data management.

Categories and Subject Descriptors

H.2.5 [**Database Management**]: Heterogeneous Databases; H.2.4 [**Database Management**]: Systems—*Distributed databases*

General Terms

Theory

Keywords

business artifact, distributed data, probabilistic data, Web data, workflow

1. INTRODUCTION

Centralized data management has matured with relational database systems. This came from the combination and cooperation of a very active research community and a very successful industry. When needed, formal tools came very handy such as first-order-logic for specifying queries and dependencies. Sometimes, they had to be developed from scratch, e.g., query optimization or concurrency control. As a consequence, the systems are now very reliable and the corresponding science is well developed.

We are interested here in Web data management. The focus is not on harvesting huge quantities of data from the Web and then managing them in a centralized manner, on a server or in a farm of servers. We are concerned here with the management of distributed on the Web in the large scale, i.e., large volume of data on a large number of autonomous, heterogeneous systems.

For such setting, the foundations of data management such as relational algebra and calculus or concurrency control do not suffice. There are a number of reasons for that:

Trees. The exchange standard for the Web is based on data trees (HTML, XML, JSON). Foundations for tree data management are being developed, e.g., in the European FoX project.[1] Automata techniques and other logics such as monadic-second-order tend to come in to complement relational calculus and algebra. Some of the works in Webdam has been centered around Active XML (XML trees with embedded service calls).

Distribution. Data is distributed and this is being more and more considered, e.g., linked data [8]. Now, we are also interested in distributing the data processing. There are many reasons for this. One is, for instance, that one may want to keep more control over private data (unlike in centralized approaches such as actual social network systems).

Imprecision and inconsistencies. A critical dimension of the problem is the imprecise and uncertain nature of data on the Web. Furthermore, one typically finds incorrect data or may introduce errors, e.g., by wrongly interpreting someone else's ontology. There is no alternative but accepting this imprecision and the inconsistencies that arise. As a consequence, the evaluation of the quality of information and information sources is essential.

Collaboration and distributed workflows. The orchestration of distributed activities (under the responsibility of a conductor) and their choreography (when they are fully autonomous) are complex issues that are essential for a wide range of applications including notably, e-commerce systems, business processes, health-care and scientific workflows. The difficulty is to guarantee consistency or more generally, quality of service, and a challenge is to be able to statically verify critical properties of the system.

These are the main directions of the works in Webdam.

*This work has been partially funded by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013) ERC grant Webdam, agreement 226513. http://webdam.inria.fr/

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[1] http://fox7.eu/

Webdam also put a lot of effort in developing a textbook (advanced undergraduate or graduate level) on *Web data management*, published at Cambridge University Press [5]. The book is available in PDF and HTML on the Webdam Web site, `http://webdam.inria.fr/`.

The paper is organized as follows. In Section 2, we discuss the work on distributed knowledge bases. Section 3 deals with probabilistic approaches to Web data management. In Section 4, we consider the issue of workflows. The last section is a conclusion. We did not want to include an extensive list of publications. We mention a few key ones that can be used as entry points in the various facets of the work. The full list may be found on the Webdam Web site.

2. DISTRIBUTED DATA MANAGEMENT

Information of interest may be found on the Web in a variety of forms, in many systems, with different access protocols. For instance, a standard user may have information on many devices (smartphone, laptop, TV box, etc.), many systems (mailers, blogs, Web sites, etc.), many social networks (Facebook, Picasa, etc.). This same user may have access to more information from family, friends, associations, companies, etc., or organizations (tax, health, etc.).

Another dimension of the problem is distribution. Typical tasks require the collaboration of autonomous users, autonomous systems. Each actor sees only partial, incomplete information. Actors may also have imprecise, conflicting opinions. As a consequence, their collaboration entails notions such as trust and probabilities.

The control and management of this diversity are today beyond the skill of casual users. Facing similar issues, companies see the cost of managing and integrating information skyrocketing. The thesis of Webdam is that managing this diversity of data can be achieved using a distributed knowledge base handling both data and meta-data, as well as access control and localization information, in a unique holistic setting. We believe that complex Web data management tasks currently requiring deep expertise will be greatly facilitated by the automatic reasoning of the inference engine of the knowledge base. This motivated works on a Datalog-style language for a distributed setting dealing with inconsistencies and probabilities.

The management of data in a heavily distributed setting such as the Web poses a number of challenging issues including: how to find data, how to control data access, and how to integrate data coming from different sources.

A distributed knowledge base. Webdam is investigating a holistic approach to support these complex data management tasks which is based on reasoning over a distributed knowledge base. Let us consider some of the reasoning tasks that may be needed:

- A particular source may have published knowledge, say in RDF and RDFS. Reasoning is needed to query this knowledge.
- When accessing different information sources, knowledge is needed to align their concepts and relations.
- It may be necessary to use knowledge to localize data, e.g., find which systems hold some information of interest. Also, when the data is localized, some simple reasoning may be required to understand how this new source of information should be used. Today, it sometimes involves downloading some code (application). We envision that this could be achieved by obtaining some knowledge (declarative rules).
- Reasoning may be needed to manage access rights (to control who has access to some information) as well as to convince others that one does have access to some information.
- Reasoning may be needed to choose among contradicting information, to evaluate the confidence in some data or more generally the trust in some source.

WebdamExchange. We have proposed a knowledge-base model, called WebdamExchange, for sharing information on the Web, where the information is hosted on different machines that may use different access control and distribution schemes. The model uses logical statements for specifying data, access control, distribution and knowledge about other peers. The statements can be communicated, replicated, queried, and updated, while keeping track of time and provenance. This unified basis allows applications to reason about which data is accessible, where it resides, and how to retrieve it securely.

A system supporting this model has been implemented and demonstrated at the ICDE 2011 conference [7]. The demonstration illustrates how users can keep control over their data even in a social network that facilitates exchanges. In particular, it shows how users within very different data distribution schemes (centralized, DHT, gossiping in an unstructured P2P, etc.) and different access control schemes, can transparently collaborate while keeping a good control over their own data even when some of their data resides in standard website or social network systems such as Facebook. The demonstration also illustrates how users can even control data using a device with limited storage and processing capabilities such as a smartphone.

Webdamlog. In WebdamExchange, the peers' policies are hard-wired (coded in Java, in the prototype system). One would like users to be able to specify their own policies or customize existing ones. For that, we believe that the specification of modern distributed data intensive applications should be based on a reasonably simple (declarative) language that would hide most of the unnecessary details of distributed data management. This is in the spirit of work from UC Berkeley on declarative programming for distributed systems [12], notably around the Dedalus language. We see as very encouraging the performance they achieved with Datalog-style languages for applications such as Internet routing [14].

In this spirit, we introduced in [2] a novel Datalog-style rule-based language, called Webdamlog. In this language, peers exchange messages (i.e., logical facts) as well as rules. The model is formally defined, and its interest for distributed data management is illustrated through a variety of examples. We validate the semantics of our model by showing that, under certain natural conditions, our semantics converges to the same semantics as the centralized system with the same rules. Indeed, we can show this is even true when updates are considered. Another major contribution of this work is a study of the impact on expressiveness of "delegations" (the installation of rules by a peer in some other peer) and explicit timestamps. Work is continuing in this direction notably by considering inconsistencies and imprecision (based on functional dependencies and probabilities).

3. A PROBABILISTIC WORLD

The information found on the Web is typically uncertain, imprecise, possibly inconsistent. Also we may wrongly interpret it, which leads to issues such as data quality or trust. This is a major theme for Webdam. Our approach is based on probability theory and, more precisely, on probabilistic trees (i.e., probabilistic XML), since trees are seen as a natural model for Web data. We have provided solid foundations for this topic, studying the management of probabilistic XML documents. We have also developed probabilistic approaches to various practical Web data management issues: trust inference, ontology matching, summarization of XML corpora.

Probabilistic XML. We have studied in depth the problem of modeling, querying, and updating probabilistic trees that represent uncertain semi-structured information. In particular, we have proposed new concise and tractable representation systems for infinite (discrete or continuous) probability distributions over trees; we have shown how to run aggregate queries on probabilistic XML documents; we have exhibited a dichotomy between tractable and intractable queries on probabilistic XML models with local correlations that lie in the presence of joins in the query; we have shown how different probabilistic models allow different forms of tractability for updates. Rather than citing individual works, we refer to [13] for a survey on the literature on probabilistic XML, including Webdam works.

Probabilistic Approaches to Web Data Management. A number of typical problems in Web data management lend themselves naturally to probabilistic modeling, usually by assuming some form of locality of the probabilistic correlations. This assumption can be seen as a simplification that allows us to get good approximations of the general case with correlations. Webdam has applied this methodology to three different problems of practical interest:

Corroboration [11]. In an environment with many independent participants, one typically finds conflicting opinions. We have studied the problem of corroborating information coming from a large number of participants, to determine the trust in sources and the truth value of facts. We have proposed and evaluated various probabilistic algorithms that provide improvements over known techniques such as voting.

Ontology matching [16]. More and more large-scale ontologies are available on the Web, either manually created or constructed using information extraction techniques. These ontologies, that use different identifiers for the same concept or entities, need to be integrated. Previous approaches to matching ontologies either focused on schema matching, or on instance matching. We provide a novel holistic approach to ontology matching, with probabilistic foundations, that combines schema matching with instance matching. The resulting system performs very well in practice, and is applicable to the largest ontologies available on the Web.

Summarization of XML corpora [1]. In order to make sense of large corpora of XML documents, we have showed how to build an *optimal* probabilistic generator for these documents in the presence of an XML schema. The approach is extended to the important case when the schema incorporate constraints on data values. This has application in summarization, testing, or querying of XML streams.

4. SEQUENCING DISTRIBUTED TASKS

When supporting complex activities in a Web setting, one typically has to organize the cooperation between possibly many systems, and notably the sequencing of their tasks. The specification of such sequencing, sometimes referred to as choreography, is little understood. We are pursuing an original approach that models tasks with pieces of data, that are called business artifacts (following IBM terminology). The evolution of an artifact is constrained by rules on the evolution of the data. Using this approach, we are pursuing fundamental works in order to understand the intrinsic nature of workflows involving distributed systems. Webdam is also active in the area of distributed workflows for Web data management, with pioneering works and fundamental results.

There has recently been a proliferation of workflow specification frameworks, notably data-centric, in response to the need to support increasingly ubiquitous processes centered around databases. Prominent examples include e-commerce systems, enterprise business processes, health-care and scientific workflows. In contrast to traditional process-centric formalisms, data-aware formalisms treat data as first-class citizens. One such formalism, proposed at IBM by Nigam and Caswell in 2003 [15], is that of *business artifacts*. These are workflows that place data at the center of the process, controlling sequencing by constraining the evolution of the data. This is also the philosophy of the Webdam approach to workflows, where evolving trees (Active XML) are placed at the center of the workflow.

In brief, Active XML consists of XML documents (the standard format for data exchange on the Web) with embedded function calls. The state of a document evolves depending on the result of internal function calls (local computations) or external ones (interactions with users or other services). Functions can be naturally used to model tasks in a workflow. They return documents that may be active, so may in turn activate new sub-tasks, thus having the ability to naturally specify a hierarchy of tasks.

Verifying temporal properties of runs. Static analysis of complex data-centric workflows is important in order to verify critical properties, compliance to regulations, and to perform optimizations. Many verification tasks consist in checking that all runs of the workflow satisfy some desirable property (e.g., in an e-commerce application, one might wish to verify that *"no product can be delivered before it has been paid"*) The main technical challenge to static verification lies in the fact that data-centric workflows are infinite-state systems, as the domain of the data is infinite. We studied the automatic verification of temporal properties of runs of data-centric workflows, for a variant of IBM's business artifacts as well as for Active XML workflows.

For business artifacts, we consider an abstraction of the IBM model, where the data consists of records of data values that evolve under rules that query an underlying database. We build upon the work of [10], which identifies a restricted class of artifacts for which verification is feasible. However, these early results suffer from an important limitation: they fail in the presence of even very simple data dependencies or arithmetic, both crucial to real-life business processes. In [9],

we extend the artifact model and verification results to alleviate this limitation. We identify a practically significant class of business artifacts with data dependencies and arithmetic, for which verification is decidable. This work was carried out in collaboration with UC San Diego and IBM.

For Active XML workflows, we study in [6] the verification of temporal properties of runs specified in a tree-pattern-based temporal logic, namely Tree-LTL, expressing a rich class of semantic properties of the application. The main results establish the boundary of decidability and the complexity of automatic verification of Tree-LTL properties.

Comparing workflow specification frameworks. There is an increasing diversity of competing approaches to specifying data-centric workflows, that differ in the data model and the control mechanism for the sequencing of tasks. For example, the control for Active XML workflows can be specified by automata, pre-and-post conditions for function calls, or temporal logic formulas. Comparing workflow specification formalisms is intrinsically difficult because of the lack of a standard yardstick for expressiveness. In recent work [4], we develop a flexible approach for comparing workflow specification languages, in which the pertinent aspects to be taken into account are defined by *views*. We use views to compare the expressiveness of different workflow specification mechanisms. For example, we show that the different control mechanisms for Active XML workflows are largely equivalent, an indication of the robustness of the model.

The Active XML Artifact model. In [3], we introduce the Active XML artifact model, a variant of Active XML tailored to capturing data and workflow management activities in distributed settings. We argue that the model is a natural extension of the business artifact model of [15]. A prototype named AXART based on this model has been developed and demonstrated at VLDB 2010. It uses an application taken from the movie industry, that specifies task sequencing when managing actors applications for roles in films. Because the system builds upon Active XML, it allows for complex organization of tasks. It supports different ways of expressing workflow constraints, a rather unique feature.

5. CONCLUSION

We summarized some of the contributions of Webdam that are most central to the project. Due to space limitations, we did not discuss a number of other interesting works, such as ontology watermarking or Web archiving. The focus of the project is on foundations and theory. However, we also work on systems issues in order to validate some of the concepts. For instance, we are developing a Webdamlog engine and a probabilistic XML query system.

Clearly, developing a formal model for Web data management is extremely challenging. Because of the complexity of the issues, we started by tackling different facets of the problem in isolation, and then moved to more unified approaches. For instance, to address the problem of imprecision, we used a probabilistic XML model, whereas distributed knowledge bases were studied using precise data. We recently started unifying both approaches, by extending the probabilistic approach to distributed knowledge bases. Altogether, we believe we are making progress on the different fronts and in the general understanding of the problem.

6. REFERENCES

[1] S. Abiteboul, Y. Amsterdamer, D. Deutch, T. Milo, and P. Senellart. Finding optimal probabilistic generators for XML collections. In *Proceedings of the International Conference on Database Theory*, 2012.

[2] S. Abiteboul, M. Bienvenu, A. Galland, and E. Antoine. A rule-based language for Web data management. In *Proceedings of the Symposium on Principles of Database Systems*, 2011.

[3] S. Abiteboul, P. Bourhis, A. Galland, and B. Marinoiu. The AXML artifact model. In *Proceedings of the International Conference on Temporal Representation and Reasoning*, 2009.

[4] S. Abiteboul, P. Bourhis, and V. Vianu. Comparing workflow specification languages: a matter of views. In *Proceedings of the International Conference on Database Theory*, 2011.

[5] S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset, and P. Senellart. *Web Data Management*. Cambridge University Press, 2012. Available online at `http://webdam.inria.fr/Jorge/`.

[6] S. Abiteboul, L. Segoufin, and V. Vianu. Static analysis of active XML services. *ACM Transactions on Database Systems*, 34(4), 2009.

[7] E. Antoine, A. Galland, K. Lyngbaek, A. Marian, and N. Polyzotis. Social networking on top of the webdamexchange system. In *Proceedings of the International Conference on Data Engineering*, 2011.

[8] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 2009.

[9] E. Damaggio, A. Deutsch, and V. Vianu. Artifact systems with data dependencies and arithmetic. In *Proceedings of the International Conference on Database Theory*, 2011.

[10] A. Deutsch, R. Hull, F. Patrizi, and V. Vianu. Automatic verification of data-centric business processes. In *Proceedings of the International Conference on Database Theory*, 2009.

[11] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating Information from Disagreeing Views. In *Proceedings of the International Conference on Web Search and Data Mining*, 2010.

[12] J. M. Hellerstein. Datalog redux: experience and conjecture. In *Proceedings of the Symposium on Principles of Database Systems*, 2010.

[13] B. Kimelfeld and P. Senellart. Probabilistic XML: Models and complexity, 2011. Preprint available at `http://pierre.senellart.com/publications/kimelfeld2012probabilistic.pdf`.

[14] B. T. Loo, T. Condie, M. N. Garofalakis, D. E. Gay, J. M. Hellerstein, P. Maniatis, R. Ramakrishnan, T. Roscoe, and I. Stoica. Declarative networking. *Communications of the ACM*, 52(11), 2009.

[15] A. Nigam and N. S. Caswell. Business artifacts: An approach to operational specification. *IBM Systems Journal*, 42(3), 2003.

[16] F. M. Suchanek, S. Abiteboul, and P. Senellart. PARIS: Probabilistic alignment of relations, instances, and schema. *Proceedings of the VLDB Endowment*, 5(3), 2011.