Path queries for Web data

Title

Path queries for Web data

Supervisors

Cristina Sirangelo
Tél: 01 47 40 77 86
Web: `http://sirangelo.info`
Email: `cristina.sirangelo@lsv.ens-cachan.fr`
Laboratoire Spécification et Vérification (LSV)
École Normale Supérieure de Cachan
61, avenue du Président Wilson
94235 Cachan CEDEX

Benoît Groz & Nicole Bidoit
Web: `https://www.lri.fr/~groz/` & Web: `https://www.lri.fr/~bidoit/`
Email: `Benoit.Groz@lri.fr` & `nicole.bidoit@lri.fr`
Laboratoire de Recherche en Informatique (LRI)
Université Paris Sud 11
Bâtiment 650 (PCRI),
91405 Orsay Cedex, France

Description

Context. Traditional databases, ubiquitous in transactional applications, websites etc. follow the relational model, representing data with a fixed structure. The recent spread of graph-structured data such as *linked open data* and social networks motivates the development of new data models, more suitable to today's Web applications. *Path queries* are a logical formalism providing a fundamental mechanism for investigating the structure of graph data. Unlike basic SQL queries over relational data, path queries allow to express reachability properties between graph nodes. Recent developments have targeted the extension of path queries with new features allowing to query not only the graph structure, but also the actual data associated to graph nodes [LMV13].

We plan to investigate techniques for the efficient evaluation of queries expressed in these new languages.

Scientific goals. This PhD proposal intends to investigate the following aspects:

1) Under which conditions one can exploit the result of some pre-computed path queries (called the views) to answer a new query? The problem of answering queries using views is relevant in many contexts including data integration and query optimization [Hal00, NSV10, CDGLV00], but many questions remain open in many basic settings. Moreover path queries pose new challenges, given their ability to express a limited form of recursion. We intend to study the possibility of rewriting path queries over views using fragments of *Datalog*, a fixed-point logic with good computational properties. Our starting point will be the most recent results about Datalog rewritability of regular path queries [FSS14]. We then plan to apply results to a relevant application of query rewriting over views, namely *scale-independent data access* [ALK+13, FGL14]. In this context we are interested in finding conditions guaranteeing that a query can be answered using a bounded amount of data, and views can be used to this purpose.

2) Given that queries may return a large number of answers, is it possible to compute a first set of answers, and then efficiently enumerate the remaining ones? The starting point will be to investigate whether techniques developed in the context of relational and tree-structured data [DSS14, LM14] can be adapted to path queries with data operators.

3) How can queries be evaluated over rapidly evolving and streaming data? In particular, in response to data updates, how can queries be evaluated incrementally? In the streaming setting the graph is not entirely stored, but progressively revealed in one or several passes; in which cases queries can still be answered correctly? Similar questions have been investigated for XML data [KMV07, SS07, GN11]. Over graphs it is natural to expect that multi-pass streaming algorithms are more meaningful than in the context of XML data transfer (the graph, although distributed, may be considered available for crawling).

Another possible starting point are streaming graph algorithms [McG14]. These are usually concerned with the graph structure only, while database queries we are interested in investigating both data and structure.

We plan to validate results trough experiments.

Positioning This research will be conducted within two teams: DAHU at LSV and LaHDAK at LRI, having high expertise on the main topics of this proposal, including regular path queries [BBG13], query rewriting [FSS14, GSC+14], streaming query evaluation [SS07], and enumeration [DSS14].

Graph queries, incremental and streaming query evaluation are also investigated by teams with which the supervisors regularly collaborate (L.Libkin from the university of Edinburgh, and the LINKS team at Lille).

Required background The PhD candidate is required to have a background in logic and databases. Programming skills are also welcome.

References

[ALK+13] Michael Armbrust, Eric Liang, Tim Kraska, Armando Fox, Michael J. Franklin, and David A. Patterson. Generalized scale independence through incremental precomputation. In *ACM Intl. Conference on Management of Data (SIGMOD)*, pages 625–636, 2013.

[BBG13] Guillaume Bagan, Angela Bonifati, and Benoit Groz. A trichotomy for regular simple path queries on graphs. In *ACM Symp. on Principles of Database Systems (PODS)*, PODS '13, pages 261–272. ACM, 2013.

[CDGLV00] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Moshe Y. Vardi. View-based query processing and constraint satisfaction. In *ACM/IEEE Symp. on Logic in Computer Science (LICS)*, pages 361–371. IEEE, 2000.

[DSS14] Arnaud Durand, Nicole Schweikardt, and Luc Segoufin. Enumerating answers to first-order queries over databases of low degree. In *ACM Symp. on Principles of Database Systems (PODS)*, pages 121–131, 2014.

[FGL14] Wenfei Fan, Floris Geerts, and Leonid Libkin. On scale independence for querying big data. In *ACM Symp. on Principles of Database Systems (PODS)*, pages 51–62, 2014.

[FSS14] Nadime Francis, Luc Segoufin, and Cristina Sirangelo. Datalog rewritings of regular path queries using views. In *Intl. Conf. on Database Theory (ICDT)*, pages 107–118, 2014.

[GN11] Olivier Gauwin and Joachim Niehren. Streamable fragments of forward xpath. In *Implementation and Application of Automata - 16th International Conference, CIAA 2011*, pages 3–15, 2011.

[GSC+14] Benoît Groz, Slawomir Staworko, Anne-Cécile Caron, Yves Roos, and Sophie Tison. Static analysis of XML security views and query rewriting. *Inf. Comput.*, 238:2–29, 2014.

[Hal00] Alon Halevy. Theory of answering queries using views. *SIGMOD Record*, 29(1):40–47, 2000.

[KMV07] Viraj Kumar, P. Madhusudan, and Mahesh Viswanathan. Visibly pushdown automata for streaming xml. In *Proceedings of*

the 16th International Conference on World Wide Web, WWW '07, pages 1053–1062. ACM, 2007.

[LM14] Katja Losemann and Wim Martens. MSO queries on trees: enumerating answers under updates. In *Joint Meeting CSL-LICS '14, Vienna, Austria, 2014*, page 67, 2014.

[LMV13] Leonid Libkin, Wim Martens, and Domagoj Vrgoč. Querying graph databases with xpath. In *Intl. Conf. on Database Theory (ICDT)*, 2013.

[McG14] Andrew McGregor. Graph stream algorithms: A survey. *SIGMOD Rec.*, 43(1):9–20, May 2014.

[NSV10] Alan Nash, Luc Segoufin, and Victor Vianu. Views and queries: Determinacy and rewriting. *ACM Transactions on Database Systems*, 35(3), 2010.

[SS07] Luc Segoufin and Cristina Sirangelo. Constant-memory validation of streaming XML documents against dtds. In *Intl. Conf. on Database Theory (ICDT)*, pages 299–313, 2007.