Ersch. in: Service-oriented computing : 8th International Conference, ICSOC 2010, San Francisco, CA, USA, December 7 - 10, 2010; proceedings / Paul P. Maglio ...(eds.). - Berlin [u.a.] : Springer, 2010. - S. 715-716. - (Lecture notes in computer science ; 6470 : Service science). - ISBN 978-3-642-17357-8

Panta Rhei: Optimized and Ranked Data Processing over Heterogeneous Sources

Daniele Braga, Francesco Corcoglioniti,
Michael Grossniklaus, and Salvatore Vadacca

Dipartimento di Elettronica e Informazione - Politecnico di Milano
Via Ponzio, 34/5 - 20133 Milano, Italy
{braga,corcoglioniti,grossniklaus,vadacca}@elet.polimi.it

Abstract. In the era of digital information, the value of data resides not only in its volume and quality, but also in the additional information that can be inferred from the combination (aggregation, comparison and join) of such data. There is a concrete need for data processing solutions that combine distributed and heterogeneous data sources, such as Web services, relational databases, and even search engines, that can all be modeled as services. In this demonstration, we show how our *Panta Rhei* model addresses the challenge of processing data over heterogeneous sources to provide feasible and ranked combinations of these services.

Keywords: Panta Rhei, data processing, search computing.

1 Introduction and Contributions

Users have the need to express complex queries over multiple sources and heterogeneous domains. We define *search computing systems* [1] as a new class of systems aimed at responding to multi-domain queries, i.e., queries over multiple semantic fields of interest.

Data sources currently available on the Web share some common features: results are brought to the user as small chunks of information, data is retrieved in a given order and more or less sophisticated interfaces allow querying the data. We model these sources as Web "services" to be composed in order to solve a multi-domain query. Currently existing solutions to the problem, such as BPEL, lack in real-time adaptivity to failure and on-line statistics, optimization and performance of the process, and definition of a "global" ranking.

These requirements led us to the definition of *Panta Rhei*, an execution model supporting multi-domain queries, whose contribution is three-fold.

- a *physical query algebra*, representing query plans as a workflow (comprising both data and control) where the basic node is the service implementation and composition of nodes is performed according to a set of simple and recursive production rules (parallel and pipe join of nodes and concatenation of modifiers, such as selection, projection and sorting).
- an *optimization layer*, choosing a feasible plan topology and defining the optimal parameters based on off-line service statistics.

– a *runtime environment*, executing plans expressed by means of the physical algebra and adapting the execution according to possible failure events and on-line service statistics.

A query workbench has been deployed to highlight internal details of the model and to ease debugging and testing of optimization heuristics.

2 Demonstration Storyboard

Our demonstration highlights the four steps of the query process. The conceptual aspects our model will be presented with the joint use of slides, whereas the practical aspects will be showcased by means of a query execution workbench. To be more precise, the demonstration is organized as follows:

1. *Introduction.* Introduction of the *Panta Rhei* model, its conceptual background and novelty of our approach.
2. *Query design.* Specification of the query in a Datalog-like conjunctive form and definition of the optimization parameters to be applied. We will show how different optimization parameters and statistics lead to different topologies.
3. *Logical query plan.* Specification of a workflow with quantitative estimates of the size of partial results determined by the planner which exploits the available degrees of freedom to fix the topology, the number and sequence of service invocations, the join strategies, etc.
4. *Physical query plan.* Definition of both data and control flow in the execution of a query. The execution engine instantiates physical operators (units) of the compiled plan and the query is orchestrated accordingly.
5. *Query execution.* Execution of the physical plan and specification of the input attributes. Plan execution can be constrained in terms of number of commands, resulting combinations or execution time.
6. *Conclusion and future work.* Summary and possible future directions.

A preview of the demo can be found at *http://www.search-computing.it/demo/qp*.

Acknowledgments. This research is part of the Search Computing (SeCo) project [*www.search-computing.it*], funded by the European Research Council (ERC). We thank all the members of the SeCo team for their contributions.

Reference

1. Ceri, S., Brambilla, M. (eds.): Search Computing. LNCS, vol. 5950. Springer, Heidelberg (2010)