IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université de Grenoble Alpes

Activity Report 2018

Project-Team TYREX

Types and Reasoning for the Web

IN COLLABORATION WITH: Laboratoire d'Informatique de Grenoble (LIG)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team TYREX

Creation of the Team: 2012 November 01, updated into Project-Team: 2014 July 01

Keywords:

Computer Science and Digital Science:

A2.1.1. - Semantics of programming languages
A2.1.4. - Functional programming
A2.1.7. - Distributed programming
A2.1.10. - Domain-specific languages
A2.2.1. - Static analysis
A2.2.4. - Parallel architectures
A2.2.8. - Code generation
A2.4. - Formal method for verification, reliability, certification
A3.1. - Data
A3.1.1. - Modeling, representation
A3.1.2. - Data management, quering and storage
A3.1.3. - Distributed data
A3.1.6. - Query optimization
A3.1.9. - Database
A3.1.10. - Heterogeneous data
A3.1.11. - Structured data
A3.2.1. - Knowledge bases
A3.2.2. - Knowledge extraction, cleaning
A3.2.6. - Linked data
A3.3.3. - Big data analysis
A3.4. - Machine learning and statistics
A3.4.1. - Supervised learning
A5.6. - Virtual reality, augmented reality
A6.3.3. - Data processing
A7. - Theory of computation
A7.1. - Algorithms
A7.2. - Logic in Computer Science
A9.1. - Knowledge
A9.2. - Machine learning
A9.7. - AI algorithmics
A9.8. - Reasoning

Other Research Topics and Application Domains:

B6.1. - Software industry
B6.3.1. - Web
B6.5. - Information systems
B8.2. - Connected city
B9.5.1. - Computer science

B9.5.6. - Data science
B9.7.2. - Open data
B9.11. - Risk management
B9.11.2. - Financial risks

1. Team, Visitors, External Collaborators

Research Scientists
Pierre Genevès [Team leader, CNRS, Researcher, HDR]
Nabil Layaïda [Inria, Senior Researcher, HDR]

Faculty Members
Angela Bonifati [Univ de Claude Bernard, Professor, from Sep 2018, HDR]
Nils Gesbert [Institut polytechnique de Grenoble, Associate Professor]
Cécile Roisin [Univ Pierre Mendès France, Professor, HDR]

Post-Doctoral Fellow
Thibaud Michel [Univ Grenoble Alpes, until May 2018]

PhD Students
Fateh Boulmaiz [Société privée Qualifret]
Sarah Chlyah [Inria, from Mar 2018]
Amela Fejza [Univ Grenoble Alpes, from Oct 2018]
Louis Jachiet [Ecole Normale Supérieure Paris, until Aug 2018]
Muideen Lawal [Univ Grenoble Alpes]

Technical staff
Thomas Calmant [Inria]
Thibaud Michel [Inria, from Jun 2018]

Interns
Amela Fejza [Inria, from Feb 2018 until Jul 2018]
Rosa Mercedes Orihuela [Inria, until Jun 2018]

Administrative Assistant
Helen Pouchot-Rouge-Blanc [Inria]

2. Overall Objectives

2.1. Objectives

We work on the foundations of the next generation of data analytics and data-centric programming systems. These systems extend ideas from programming languages, artificial intelligence, data management systems, and theory. Data-intensive applications are increasingly more demanding in sophisticated algorithms to represent, store, query, process, analyse and interpret data. We build and study data-centric programming methods and systems at the core of artificial intelligence applications. Challenges include the robust and efficient processing of large amounts of structured, heterogeneous, and distributed data.

On the data-intensive application side, our current focus is on building efficient and scalable analytics systems. Our technical contributions particularly focus on the optimization, compilation, and synthesis of information extraction and analytics code, in particular with large amounts of data.

On the theoretical side, we develop the foundations of data-centric systems and analytics engines with a particular focus on the analysis and typing of data manipulations. We focus in particular on the foundations of programming with distributed data collections. We also study the algebraic and logical foundations of query languages, for their analysis and their evaluation.

3. Research Program

3.1. Foundations for Data Manipulation Analysis: Logics and Type Systems

We develop methods for the static analysis of queries based on logical decision procedures. Static analysis can be used to optimize runtime performance by compile-time automated modification of the code. For example, queries can be substituted by more efficient — yet equivalent — variants. The query containment problem has been a central point of research for major query languages due to its vital role in query optimization. Query containment is defined as determining if the result of one query is included in the result of another one for any dataset. We explore techniques for deciding query containment for expressive languages for querying richly structured data such as knowledge graphs. One major scientific difficulty here consists in dealing with problems close to the frontier of decidability, and therefore in finding useful trade-offs between programming expressivity, complexity, succinctness, algorithmic techniques and effective implementations. We also investigate type systems and type-checking methods for the analysis of the manipulations of structured data.

3.2. Algebraic Foundations for Query Optimization and Code Synthesis

We consider intermediate languages based on algebraic foundations for the representation, characterization, transformations and compilation of queries. We investigate extensions of the relational algebra for optimizing expressive queries, and in particular recursive queries. We explore monads and in particular monad comprehensions and monoid calculus for the generation of efficient and scalable code on big data frameworks. When transforming and optimizing algebraic terms, we rely on cost-based searches of equivalent terms. We thus develop cost models whose purpose is to estimate the time, space and network costs of query evaluation. One difficulty is to estimate these costs in architectures where data and computations are distributed, and where the modeling of data transfers is essential.

4. Application Domains

4.1. Querying Large Graphs

Increasingly large amounts of graph-structured data become available. The methods we develop apply for the efficient evaluation of graph queries over large — and potentially distributed — graphs. In particular, we consider the SPARQL query language, which is the standard language for querying graphs structured in the Resource Description Format (RDF). We also consider other increasingly popular graph query languages such as Cypher queries for extracting information from property graphs.

We compile graph queries into lower-level distributed primitives found in big data frameworks such as Apache Spark, Flink, etc. Applications of graph querying are ubiquitous and include: large knowledge bases, social networks, road networks, trust networks and fraud detection for cryptocurrencies, publications graphs, web graphs, recommenders, etc.

4.2. Predictive Analytics for Healthcare

One major expectation of data science in healthcare is the ability to leverage on digitized health information and computer systems to better apprehend and improve care. The availability of large amounts of clinical data and in particular electronic health records opens the way to the development of quantitative models for patients that can be used to predict health status, as well as to help prevent disease and adverse effects.

In collaboration with the CHU Grenoble, we explore solutions to the problem of predicting important clinical outcomes such as patient mortality, based on clinical data. This raises many challenges including dealing with the very high number of potential predictor variables and very resource-consuming data preparation stages.

4.3. Mobile and Augmented Reality Applications

The term Augmented Environments refers collectively to ubiquitous computing, context-aware computing, and intelligent environments. The goal of our research on these environments is to introduce personal Augmented Reality (AR) devices, taking advantage of their embedded sensors. These environments offer the possibility of using ubiquitous computation, communication, and sensing to enable the presentation of context-sensitive information and services to the user. AR applications often rely on 3D content and employ specialized hardware and computer vision techniques for both tracking and scene reconstruction and exploration. Our approach tries to seek a balance between these traditional AR contexts and what has come to be known as mobile AR browsing, based for instance on attitude estimation.

5. New Software and Platforms

5.1. SPARQLGX

KEYWORDS: RDF - SPARQL - Distributed computing
SCIENTIFIC DESCRIPTION: SPARQL is the W3C standard query language for querying data expressed in RDF (Resource Description Framework). The increasing amounts of RDF data available raise a major need and research interest in building efficient and scalable distributed SPARQL query evaluators.

In this context, we propose and share SPARQLGX: our implementation of a distributed RDF datastore based on Apache Spark. SPARQLGX is designed to leverage existing Hadoop infrastructures for evaluating SPARQL queries. SPARQLGX relies on a translation of SPARQL queries into executable Spark code that adopts evaluation strategies according to (1) the storage method used and (2) statistics on data. Using a simple design, SPARQLGX already represents an interesting alternative in several scenarios.
FUNCTIONAL DESCRIPTION: This software system is an implementation of a distributed evaluator of SPARQL queries. It makes it possible to evaluate SPARQL queries on billions of triples distributed across multiple nodes in a cluster, while providing attractive performance figures.

RELEASE FUNCTIONAL DESCRIPTION: - Faster load routine which widely improves this phase perfomances by reading once the initial triple file and by partitioning data in the same time into the correct predicate files. - Improving the generated Scala-code of the translation process with mapValues. This technic allows not to break the partitioning of KeyValueRDD while applying transformations to the values instead of the traditional map that was done prior. - Merging and cleaning several scripts in bin/ such as for example sgx-eval.sh and sde-eval.sh - Improving the compilation process of compile.sh - Cleaner test scripts in tests/ - Offering the possibility of an easier deployment using Docker.

- Participants: Damien Graux, Thomas Calmant, Louis Jachiet, Nabil Layaïda and Pierre Genevès
- Contact: Pierre Genevès
- Publications: Optimizing SPARQL query evaluation with a worst-case cardinality estimation based on statistics on the data - The SPARQLGX System for Distributed Evaluation of SPARQL Queries
- URL: https://github.com/tyrex-team/sparqlgx

5.2. musparql

KEYWORDS: SPARQL - RDF - Property paths
FUNCTIONAL DESCRIPTION: reads a SPARQL request and translates it into an internal algebra. Rewrites the resulting term into many equivalent versions, then choses one of them and executes it on a graph.

- Participant: Louis Jachiet
- Contact: Nabil Layaïda
- Publication: Extending the SPARQL Algebra for the optimization of Property Paths
- URL: https://gitlab.inria.fr/tyrex/musparql

5.3. MRB

Mixed Reality Browser
KEYWORDS: Augmented reality - Geolocation - Indoor geolocalisation - Smartphone
FUNCTIONAL DESCRIPTION: MRB displays PoI (Point of Interest) content remotely through panoramics with spatialized audio, or on-site by walking to the corresponding place, it can be used for indoor-outdoor navigation, with assistive audio technology for the visually impaired. It is the only browser of geolocalized data to use XML as a native format for PoIs, panoramics, 3D audio and to rely on HTML5 both for the iconic and full information content of PoIs. Positioning in MRB is based on a PDR library, written in C++ and Java and developed by the team, which provides the user's location in real time based on the interpretation of sensors. Three main modules have been designed to build this positioning system: (i) a pedometer that estimates the distance the user has walked and his speed, (ii) a motion manager that enables data set recording and simulation but also the creation of virtual sensors or filters (e.g gyroscope drift compensation, linear acceleration, altimeter), and (iii) a map-matching algorithm that provides a new location based on a given OpenStreetMap file description and the current user's trajectory.

- Participant: Thibaud Michel
- Contact: Nabil Layaïda
- Publications: On Mobile Augmented Reality Applications based on Geolocation - Attitude Estimation for Indoor Navigation and Augmented Reality with Smartphones
- URL: http://tyrex.inria.fr/projects/mrb.html

5.4. Benchmarks Attitude Smartphones

KEYWORDS: Experimentation - Motion analysis - Sensors - Performance analysis - Smartphone
SCIENTIFIC DESCRIPTION: We investigate the precision of attitude estimation algorithms in the particular context of pedestrian navigation with commodity smartphones and their inertial/magnetic sensors. We report on an extensive comparison and experimental analysis of existing algorithms. We focus on typical motions of smartphones when carried by pedestrians. We use a precise ground truth obtained from a motion capture system. We test state-of-the-art attitude estimation techniques with several smartphones, in the presence of magnetic perturbations typically found in buildings. We discuss the obtained results, analyze advantages and limits of current technologies for attitude estimation in this context. Furthermore, we propose a new technique for limiting the impact of magnetic perturbations with any attitude estimation algorithm used in this context. We show how our technique compares and improves over previous works.

- Participants: Hassen Fourati, Nabil Layaïda, Pierre Genevès and Thibaud Michel
- Partner: GIPSA-Lab
- Contact: Pierre Genevès
- URL: http://tyrex.inria.fr/mobile/benchmarks-attitude/

5.5. MedAnalytics

KEYWORDS: Big data - Predictive analytics - Distributed systems
FUNCTIONAL DESCRIPTION: We implemented a method for the automatic detection of at-risk profiles based on a fine-grained analysis of prescription data at the time of admission. The system relies on an optimized distributed architecture adapted for processing very large volumes of medical records and clinical data. We conducted practical experiments with real data of millions of patients and hundreds of hospitals. We demonstrated how the various perspectives of big data improve the detection of at-risk patients, making it possible to construct predictive models that benefit from volume and variety. This prototype implementation is described in the 2017 preprint available at: https://hal.inria.fr/hal-01517087/document.

- Participants: Pierre Genevès and Thomas Calmant
- Partner: CHU Grenoble
- Contact: Pierre Genevès
- Publication: Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission

5.6. MuIR

Mu Intermediate Representation
KEYWORDS: Optimizing compiler - Querying
FUNCTIONAL DESCRIPTION: This is a prototype of an intermediate language representation, i.e. an implementation of algebraic terms, rewrite rules, query plans, cost model, query optimizer, and query evaluators (including a distributed evaluator of algebraic terms using Apache Spark).

- Contact: Pierre Genevès

6. New Results

6.1. On the Optimization of Recursive Relational Queries

Graph databases have received a lot of attention recently as they are particularly useful in many applications such as social networks or for the semantic web. Various languages have emerged to query such graph databases. At the heart of many of those query languages, there is a construction to navigate through the graph which allows some form of recursion.The relational model has benefited from a huge body of research in the last half century and that is why many graph databases either rely on, or have adopted the techniques of, relational-based query engines. Since its introduction, the relational model has seen various attempts to extend it with recursion and it is now possible to use recursion in several SQL- or Datalog-based database systems. The optimization of recursive queries remains, however, a challenge. In this work, we introduce μ-RA, a variation of the Relational Algebra that allows for the expression of relational queries with recursion. μ-RA can express unions of conjunctive regular path queries as well as certain non-regular properties. We present its syntax, semantics and the rewriting rules we specifically devised to tackle the optimization of recursive queries. A prototype evaluator implementing these rewriting rules is shown to be more efficient than previous approaches.

These results were presented at the BDA 2018 conference [14].

6.2. A Multi-Criteria Experimental Ranking of Distributed SPARQL Evaluators

SPARQL is the standard language for querying RDF data. There exists a variety of SPARQL query evaluation systems implementing different architectures for the distribution of data and computations. Differences in architectures coupled with specific optimizations, for e.g. preprocessing and indexing, make these systems incomparable from a purely theoretical perspective. This results in many implementations solving the SPARQL query evaluation problem while exhibiting very different behaviours, not all of them being adapted to any context. We provide a new perspective on distributed SPARQL evaluators, based on multi-criteria experimental rankings. Our suggested set of 5 features (namely velocity, immediacy, dynamicity, parsimony, and resiliency) provides a more comprehensive description of the behaviours of distributed evaluators when compared to traditional runtime performance metrics. We show how these features help in more accurately evaluating to which extent a given system is appropriate for a given use case. For this purpose, we systematically benchmarked a panel of 10 state-of-the-art implementations. We ranked them using a reading grid that helps in pinpointing the advantages and limitations of current technologies for the distributed evaluation of SPARQL queries.

These results were presented at the IEEE Big Data 2018 conference [13].

6.3. SPARQL Query Containment under Schema

Query containment is defined as the problem of determining if the result of a query is included in the result of another query for any dataset. It has major applications in query optimization and knowledge base verification. The main objective of this work is to provide sound and complete procedures to determine containment of SPARQL queries under expressive description logic schema axioms. Beyond that, these procedures are experimentally evaluated. To date, testing query containment has been performed using different techniques: containment mapping, canonical databases, automata theory techniques and through a reduction to the validity problem in logic. In this work, we use the latter technique to test containment of SPARQL queries using an expressive modal logic called μ-calculus. For that purpose, we define an RDF graph encoding as a transition system which preserves its characteristics. In addition, queries and schema axioms are encoded as μ-calculus formulae. Thereby, query containment can be reduced to testing validity in the logic. We identify various fragments of SPARQL and description logic schema languages for which containment is decidable. Additionally, we provide theoretically and experimentally proven procedures to check containment of these decidable fragments. Finally, we propose a benchmark for containment solvers which is used to test and compare the current state-of-the-art containment solvers.

These results were published in the Journal on Data Semantics [4].

6.4. Selectivity Estimation for SPARQL Triple Patterns with Shape Expressions

ShEx (Shape Expressions) is a language for expressing constraints on RDF graphs. In this work we optimize the evaluation of conjunctive SPARQL queries, on RDF graphs, by taking advantage of ShEx constraints. Our optimization is based on computing and assigning ranks to query triple patterns, dictating their order of execution. We first define a set of well-formed ShEx schemas that possess interesting characteristics for SPARQL query optimization. We then define our optimization method by exploiting information extracted from a ShEx schema. We finally report on evaluation results performed showing the advantages of applying our optimization on the top of an existing state-of-the-art query evaluation system.

These results were presented at the 2018 International Conference on Web Engineering [9].

6.5. Evaluation of Query Transformations without Data

Query transformations are ubiquitous in semantic web query processing. For any situation in which transformations are not proved correct by construction, the quality of these transformations has to be evaluated. Usual evaluation measures are either overly syntactic and not very informative — the result being: correct or incorrect — or dependent from the evaluation sources. Moreover, both approaches do not necessarily yield the same result. We suggest that grounding the evaluation on query containment allows for a data-independent evaluation that is more informative than the usual syntactic evaluation. In addition, such evaluation modalities may take into account ontologies, alignments or different query languages as soon as they are relevant to query evaluation.

These results were presented at a workshop of the 2018 International Conference on World Wide Web [10].

6.6. Graph Queries: From Theory to Practice

In this work, we review various graph query language fragments that are both theoretically tractable and practically relevant. We focus on the most expressive one that retains these properties and use it as a stepping stone to examine the underpinnings of graph query evaluation along graph view maintenance. Further broadening the scope of the discussion, we then consider alternative processing techniques for graph queries, based on graph summarization and path query learning. We conclude by pinpointing the open research directions in this emerging area. These results were published in Sigmod Record Journal [3].

6.7. Query-based Linked Data Anonymization

In this work, we introduce and develop a declarative framework for privacy-preserving Linked Data publishing in which privacy and utility policies are specified as SPARQL queries. Our approach is data independent and leads to inspect only the privacy and utility policies in order to determine the sequence of anonymization operations applicable to any graph instance for satisfying the policies. We prove the soundness of our algorithms and gauge their performance through experiments.

These results were presented in the International Semantic Web Conference (ISWC 2018) [11].

6.8. Querying Graphs

Graph data modeling and querying arises in many practical application domains such as social and biological networks where the primary focus is on concepts and their relationships and the rich patterns in these complex webs of interconnectivity. In this book, we present a concise unified view on the basic challenges which arise over the complete life cycle of formulating and processing queries on graph databases. To that purpose, we present all major concepts relevant to this life cycle, formulated in terms of a common and unifying ground: the property graph data model — the predominant data model adopted by modern graph database systems.

In this book [17], we aim especially to give a coherent and in-depth perspective on current graph querying and an outlook for future developments. Our presentation is self-contained, covering the relevant topics from: graph data models, graph query languages and graph query specification, graph constraints, and graph query processing. We conclude by indicating major open research challenges towards the next generation of graph data management systems.

6.9. Backward Type Inference for XML Queries

Although XQuery is a statically typed, functional query language for XML data, some of its features such as upward and horizontal XPath axes are typed imprecisely. The main reason is that while the XQuery data model allows to navigate upwards and between siblings from a given XML node, the type model, e.g., regular tree types, can describe only the subtree structure of the given node. Recently, Giuseppe Castagna and our team independently proposed in 2015 a precise forward type inference system for XQuery using an extended type language that can describe not only a given XML node but also its context. In this work, as a complementary method to such forward type inference systems, we propose an enhanced backward type inference system for XQuery, based on an extended type language. Results include an exact type system for XPath axes and a sound type system for XQuery expressions [19].

6.10. Scalable and Interpretable Predictive Models for Electronic Health Records

Early identification of patients at risk of developing complications during their hospital stay is currently one of the most challenging issues in healthcare. Complications include hospital-acquired infections, admissions to intensive care units, and in-hospital mortality. Being able to accurately predict the patients' outcomes is a crucial prerequisite for tailoring the care that certain patients receive, if it is believed that they will do poorly without additional intervention. We consider the problem of complication risk prediction, such as patient mortality, from the electronic health records of the patients. We study the question of making predictions on the first day at the hospital, and of making updated mortality predictions day after day during the patient's stay. We develop distributed models that are scalable and interpretable. Key insights include analysing diagnoses known at admission and drugs served, which evolve during the hospital stay. We leverage a distributed architecture to learn interpretable models from training datasets of gigantic size. We test our analyses with more than one million of patients from hundreds of hospitals, and report on the lessons learned from these experiments.

These results were presented at the 2018 International Conference on Data Science and Applications [12].

6.11. Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission

We show how the analysis of very large amounts of drug prescription data make it possible to detect, on the day of hospital admission, patients at risk of developing complications during their hospital stay. We explore, for the first time, to which extent volume and variety of big prescription data help in constructing predictive models for the automatic detection of at-risk profiles. Our methodology is designed to validate our claims that: (1) drug prescription data on the day of admission contain rich information about the patient's situation and perspectives of evolution, and (2) the various perspectives of big medical data (such as veracity, volume, variety) help in extracting this information.We build binary classification models to identify at-risk patient profiles. We use a distributed architecture to ensure scalability of model construction with large volumes of medical records and clinical data. We report on practical experiments with real data of millions of patients and hundreds of hospitals. We demonstrate how the fine-grained analysis of such big data can improve the detection of at-risk patients, making it possible to construct more accurate predictive models that significantly benefit from volume and variety, while satisfying important criteria to be deployed in hospitals.

These results were published in the Big Data Research journal [6].

6.12. ProvSQL: Provenance and Probability Management in PostgreSQL

This demonstration showcases ProvSQL, an open-source module for the PostgreSQL database management system that adds support for computation of provenance and probabilities of query results. A large range of provenance formalisms are supported, including all those captured by provenance semirings, provenance semirings with monus, as well as where-provenance. Probabilistic query evaluation is made possible through the use of knowledge compilation tools, in addition to standard approaches such as enumeration of possible worlds and Monte-Carlo sampling. ProvSQL supports a large subset of non-aggregate SQL queries.

These results were published in the PVLDB journal [8].

6.13. A Method to Quantitatively Evaluate Geo Augmented Reality Applications

We propose a method for quantitatively assessing the quality of Geo AR browsers. Our method aims at measuring the impact of attitude and position estimations on the rendering precision of virtual features. We report on lessons learned by applying our method on various AR use cases with real data. Our measurement technique allows shedding light on the limits of what can be achieved in Geo AR with current technologies. This also helps in identifying interesting perspectives for the further development of high-quality Geo AR applications.

These results were presented at the ISMAR 2018 conference [15].

6.14. Attitude Estimation for Indoor Navigation and Augmented Reality with Smartphones

We investigate the precision of attitude estimation algorithms in the particular context of pedestrian navigation with commodity smartphones and their inertial/magnetic sensors. We report on an extensive comparison and experimental analysis of existing algorithms. We focus on typical motions of smartphones when carried by pedestrians. We use a precise ground truth obtained from a motion capture system. We test state-of-the-art and built-in attitude estimation techniques with several smartphones, in the presence of magnetic perturbations typically found in buildings. We discuss the obtained results, analyze advantages and limits of current technologies for attitude estimation in this context. Furthermore, we propose a new technique for limiting the impact of magnetic perturbations with any attitude estimation algorithm used in this context. We show how our technique compares and improves over previous works. A particular attention was paid to the study of attitude estimation in the context of augmented reality motions when using smartphones.

These results were published in the Pervasive and Mobile Computing journal [7].

6.15. A Hybrid Approach for Spatio-Temporal Validation of Declarative Multimedia

Declarative multimedia documents represent the description of multimedia applications in terms of media items and relationships among them. Relationships specify how media items are dynamically arranged in time and space during runtime. Although a declarative approach usually facilitates the authoring task, authors can still make mistakes due to incorrect use of language constructs or inconsistent or missing relationships in a document. In order to properly support multimedia application authoring, it is important to provide tools with validation capabilities. Document validation can indicate possible inconsistencies in a given document to an author so that it can be revised before deployment. Although very useful, multimedia validation tools are not often provided by authoring tools. This work proposes a multimedia validation approach that relies on a formal model called Simple Hyper-media Model (SHM). SHM is used for representing a document for the purpose of validation. An SHM document is validated using a hybrid approach based on two complementary techniques. The first one captures the document's spatio-temporal layout in terms of its state throughout its execution by means of a rewrite theory, and validation is performed through model checking. The second one captures the document's layout in terms of intervals and event occurrences by means of Satisfiability Modulo Theories (SMT) formulas, and validation is performed through SMT solving. Due to different characteristics of both approaches, each validation technique complements the other in terms of expressiveness of SHM and tests to be checked. We briefly present validation tools that use our approach. They were evaluated with real NCL documents and by usability tests.

These results were published in the ACM Transactions on Multimedia Computing, Communications and Applications journal [5].

7. Partnerships and Cooperations

7.1. Regional Initiatives

Data-CILE

Title: Query Compilation

Call: Appel à projet Grenoble Innovation Recherche (AGIR-Pôle)

Duration: 2016-2018

Coordinator: Nabil Layaïda

Abstract: The goal of this project is to contribute to foundational and algorithmic challenges introduced by increasingly popular data-centric paradigms for programming on distributed architectures such as spark and the massive production of big linked open data. The focus of the project is on building robust and more efficient workflows of transformations of semantic and graph web data.

BioQurate

Title: Querying and Curating Hierarchies of Biological Graphs

Funding: Fédération Informatique de Lyon (FIL)

Duration: 2018-2020

Coordinator: Angela Bonifati

Others partners: LIP/LIRIS. The project involves a bio-computing team and a database team on a common research problem

Abstract: This project aims at leveraging graph rewriting techniques of ReGraph and graph data management techniques in order to provide a persistent, robust and scalable substrate for the construction and manipulation of hierarchies of biological graphs. Moreover, we wish to investigate whether the involved graphs need further expressive graph constraints for enforcing consistency and performing data cleansing.

7.2. National Initiatives

7.2.1. ANR

CLEAR

Title: Compilation of intermediate Languages into Efficient big dAta Runtimes

Call: Appel à projets générique 2016 défi 'Société de l'information et de la communication' – JCJC

Duration: January 2017 – September 2021

Coordinator: Pierre Genevès

See also: http://tyrex.inria.fr/clear

Abstract: This project addresses one fundamental challenge of our time: the construction of effective programming models and compilation techniques for the correct and efficient exploitation of big and linked data. We study high-level specifications of pipelines of data transformations and extraction for producing valuable knowledge from rich and heterogeneous data. We investigate how to synthesize code which is correct and optimized for execution on distributed infrastructures.

DataCert

Title: Coq deep specification of security aware data integration

Call: Appel à projets Sciences et technologies pour la confiance et la sécurité numérique

Duration: January 2016 – January 2020

Participant: Angela Bonifati

Others partners: Université Paris Sud/Laboratoire de Recherche en Informatique, Université de Lille/Centre de Recherche en Informatique, Signal et Automatique de Lille, Université de Lyon/Laboratoire d'InfoRmatique en Image et Systèmes d'information.

See also: http://datacert.lri.fr/

Abstract: This project's aim is to develop a comprehensive framework handling the fundamental problems underlying security-aware data integration and sharing, resulting in a paradigm shift in the design and implementation of security-aware data integration systems. To fill the gap between both worlds, we strongly rely on deep specifications and proven-correct software, develop formal models yielding highly reliable technology while controlling the disclosure of private or confidential information.

QualiHealth

Title: Enhancing the Quality of Health Data

Call: Appel à projets Projets de Recherche Collaborative – Entreprise (PRCE)

Duration: 2018-2022

Coordinator: Angela Bonifati

Others partners: LIMOS, Université Clermont Auvergne. LIS, Université d'Aix-Marseille. HEGP, INSERM, Paris. Inst. Cochin, INSERM, Paris. Gnubila, Argonay. The University of British Columbia, Vancouver (Canada)

Abstract: This research project is geared towards a system capable of capturing and formalizing the knowledge of data quality from domain experts, enriching the available data with this knowledge and thus exploiting this knowledge in the subsequent quality-aware medical research studies. We expect a quality-certified collection of medical and biological datasets, on which quality-certified analytical queries can be formulated. We envision the conception and implementation of a quality-aware query engine with query enrichment and answering capabilities.

To reach this ambitious objectives, the following concrete scientific goals must be fulfilled : (1) An innovative research approach, that starts from concrete datasets and expert practices and knowledge to reach formal models and theoretical solutions, will be employed to elicit innovative quality dimensions and to identify, formalize, verify and finally construct quality indicators able to capture the variety and complexity of medical data; those indicators have to be composed, normalized and aggregated when queries involve data with different granularities (e.g., accuracy indications on pieces of information at the patient level have to be composed when one queries cohort) and of different quality dimensions (e.g., mixing incomplete and inaccurate data); and (2) In turn, those complex aggregated indicators have to be used to provide new quality-driven query answering, refinement, enrichment and data analytics techniques. A key novelty of this project is the handling of data which are not rectified on the original database but sanitized in a query-driven fashion: queries will be modified, rewritten and extended to integrate quality parameters in a flexible and automatic way.

7.2.2. PERSYVAL-lab LabEx

Title: Mobile Augmented Reality Applications for Smart Cities

Call: Persyval Labex (“Laboratoire d’excellence”).

Duration: 2014 – 2018

Coordinators: Pierre Genevès and Nabil Layaïda

Others partners: NeCS team at GIPSA-Lab laboratory.

Abstract: The goal of this project is to increase the relevance and reliability of augmented reality (AR) applications, through three main objectives:

1. Finding and developing appropriate representations for describing the physical world (3D maps, indoor buildings, ways...), integrated advanced media types (3D, 3D audio, precisely geo-tagged pictures with lat., long. and orientation, video...)
2. Integrating the different abstraction levels of these data streams (ranging from sensors data to high level rich content such as 3D maps) and bridging the gap with Open Linked Data (the semantic World). This includes opening the way to query the environment (filtering), and adapt AR browsers to users’ capabilities (e.g. blind people). The objective here is to provide an open and scalable platform for mobile-based AR systems (just like the web represents).
3. Increasing the reliability and accuracy of localization technologies. Robust and high-accuracy localization technologies play a key role in AR applications. Combined with geographical data, they can also be used to identify user-activity patterns, such as walking, running or being in an elevator. The interpretation of sensor values, coupled with different walking models, allows one to ensure the continuity of the localization, both indoor and outdoor. However, dead reckoning based on Inertial Navigation Systems (INS) or Step-and-Heading Systems (SHS) is subject to cumulative errors due to many factors (sensor drift (accelerometers, gyroscopes, etc.), missed steps, bad estimation of the length of each stride, etc.). One objective is to reduce such errors by merging and mixing these approaches with various external signals such as GPS and Wi-Fi or relying on the analyses of user trajectories with the help of a structured map of the environment. Some filtering methods (Kalman Filter, observer, etc.) will be useful to achieve this task.

7.3. International Research Visitors

7.3.1. Visits of International Scientists

We had short visits from Wim Martens (University of Bayreuth, Germany) and Efthymia Tsamoura (University of Oxford, UK).

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events Organisation

8.1.1.1. Member of the Organizing Committees

- P. Genevès is member of the Organizing Committee of BDA 2019.
- A. Bonifati is a permanent member of ICDT Council (The International Conference on Database Theory).

8.1.2. Scientific Events Selection

8.1.2.1. Chair of Conference Program Committees

- A. Bonifati is Co-chair of the SIGMOD 2019 Workshops.

8.1.2.2. Member of the Conference Program Committees

- P. Genevès has been program committee member for the 27th International Joint Conference on Artificial Intelligence (IJCAI'18) and for the 23rd European Conference on Artificial Intelligence (ECAI'18) and for the 18th ACM Symposium on Document Engineering (DocEng'18).
- A. Bonifati has been program committee member of VLDB 2019, AAAI 2019, ICDE 2019, EDBT 2019, SIGMOD 2019 (senior PC member), PODS 2019, DEBS 2019, ICDT 2020.

8.1.2.3. Reviewer

- P. Genevès has been reviewer for the ICALP 2018, CHI 2018, and BDA 2018 conferences.
- C. Roisin has been reviewer for the 18th ACM Symposium on Document Engineering (DocEng'18).

8.1.3. Journal

8.1.3.1. Member of the Editorial Boards

- A. Bonifati is Associate Editor of ACM Trans. on Database Systems.
- A. Bonifati is Associate Editor of the VLDB Journal.

8.1.3.2. Reviewer - Reviewing Activities

- P. Genevès has been reviewer for the Sensors journal.

8.1.4. Invited Talks

- P. Genevès gave an invited talk for the Data Science in the Alps Workshop (March 20th, 2018): "On the Prediction of At-Risk Patient Profiles with Big Prescription Data"
- P. Genevès gave an invited seminar at the University of Fribourg (March 27th, 2018): "Queries for Trees and Graphs: Static Analysis and Code Synthesis"

8.1.5. Scientific Expertise

- C. Roisin has been designed as expert for the reviewing of a research proposal of the American University of Beirut.
- P. Genevès has been a scientific expert at ANRT for the CIFRE funding process.

8.1.6. Research Administration

- P. Genevès is responsible for the Computer Science Specialty at the Doctoral School MSTII (ED 217)
- P. Genevès has been Member of a Hiring Committee at IRIF.
- C. Roisin is a member of the CNU (Conseil National des Universités).
- C. Roisin is a member of the Inria Grenoble Inria-Hub committee.

- N. Layaïda is a member of the experts pool (selection committee) of the minalogic competitive cluster.
- A. Bonifati and N. Layaïda are members of the Scientific Board of Digital League, the digital cluster of Auvergne-Rhone-Alpes.
- A. Bonifati is coordinator of the theme « Masses de Données » at Liris and at « Fédération d'Informatique de Lyon » (FIL).

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

- Licence : C. Roisin, Programmation C, 12h eq TD, L2, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Architecture des réseaux, 112h eq TD, L1, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Services réseaux, 22h eq TD, L2, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Introduction système Linux, 21h eq TD, L1, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Système et réseaux, 14h eq TD, L3, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Tutorat pédagogique de 4 apprentis, 20h eq TD, L3, IUT2, Univ. Grenoble-Alpes
- Licence : C. Roisin, Suivi pédagogique de 20 étudiants (responsable de la Licence Professionnelle MI-ASSR), 13h eq TD, L1, IUT2, Univ. Grenoble-Alpes
- Licence : N. Gesbert, 'Logique pour l'informatique', 45 h eq TD, L3, Grenoble INP
- Licence : N. Gesbert, 'Bases de la programmation impérative', 30 h eq TD, L3, Grenoble INP
- Master : N. Gesbert, academic tutorship of an apprentice, 10 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, 'Fondements logiques pour l'informatique', 16 h 30 eq TD, M1, Grenoble INP
- Master : N. Gesbert, 'Construction d'applications Web', 21 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, 'Analyse, conception et validation de logiciels', 30 h eq TD, M1, Grenoble INP
- N. Gesbert is responsible of the L3-level course 'logique pour l'informatique' (25 apprentices) and of the M1-level course 'construction d'applications Web' (72 students).
- P. Genevès is responsible and teacher in the the M2-level course 'Semantic Web: from XML to OWL' of the MOSIG program at UGA (36h)
- P. Genevès is responsible and teacher in the the M2-level course 'Accès à l'information: du web des données au web sémantique' of the ENSIMAG ISI 3A program at Grenoble-INP (30h)

8.2.2. Supervision

- PhD: Louis Jachiet, On the foundations for the compilation of web data queries: optimization and distributed evaluation of SPARQL, University Grenoble Alpes. PhD Thesis defended on September 15th, 2018. Co-supervised by Nabil Layaïda and Pierre Genevès.
- PhD in progress: Muideen Lawal, Cost models for optimizing compilers based on mu-terms, PhD started in October 2017, supervised by Pierre Genevès.
- PhD in progress: Raouf Kerkouche, Privacy-preserving predictive analytics with big prescription data, PhD started in October 2017, co-supervised by Pierre Genevès and Claude Castelluccia.
- PhD in progress: Fateh Boulmaiz, Distributed representations of large-scale graphs, PhD started in November 2017, co-supervised by Pierre Genevès and Nabil Layaïda.
- PhD in progress: Sarah Chlyah, Algebraic foundations for the synthesis of optimized distributed code, PhD started in March 2018, supervised by Pierre Genevès.
- PhD in progress: Amela Fejza, On the extended algebraic representations for analytical workloads, PhD started in October 2018, supervised by Pierre Genevès.

8.3. Popularization

- N. Layaïda contributed to the special édition of the newspaper : Le Dauphiné libéré des enfants – C'est quoi le numérique ? – n°14 November-December 2018.
- P. Genevès presented the Tyrex team activities in 180 seconds in a dissemination event organized at Inria Montbonnot in June 2018.

8.3.1. Creation of media or tools for science outreach

- T. Michel built a mobile application for the Inria public showroom, Espace Login, called Login'AR. The application was designed to showcase the combination of indoor positioning with Augmented Reality capabilities. The application is now part of the permanent Login public exhibition. The application was developed with the help of four Grenoble INP interns (A. Convert, A. Fortecof, G. Montano and S.J. Tourangeau).

9. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] L. JACHIET. *On the foundations for the compilation of web data queries: optimization and distributed evaluation of SPARQL*, Communauté Université Grenoble Alpes, September 2018, https://hal.inria.fr/tel-01891444

[2] L. JACHIET. *On the foundations for the compilationof web data queries : optimization anddistributed evaluation of SPARQL*, Université Grenoble Alpes, September 2018, https://tel.archives-ouvertes.fr/tel-01960209

Articles in International Peer-Reviewed Journals

[3] A. BONIFATI, S. DUMBRAVA. *Graph Queries: From Theory to Practice*, in "ACM SIGMOD Record", December 2018, vol. 47, n° 4, https://hal.inria.fr/hal-01977048

[4] M. CHEKOL, J. EUZENAT, P. GENEVÈS, N. LAYAÏDA. *SPARQL Query Containment under Schema*, in "Journal on Data Semantics", April 2018, vol. 7, n° 3, pp. 133-154 [*DOI :* 10.1007/s13740-018-0087-1], https://hal.inria.fr/hal-01767887

[5] J. A. F. DOS SANTOS, D. C. MUCHALUAT-SAADE, C. ROISIN, N. LAYAÏDA. *A Hybrid Approach for Spatio-Temporal Validation of Declarative Multimedia Documents*, in "ACM Transactions on Multimedia Computing, Communications and Applications", November 2018, vol. 14, n° 4, pp. 1-24 [*DOI :* 10.1145/3267127], https://hal.inria.fr/hal-01946641

[6] P. GENEVÈS, T. CALMANT, N. LAYAÏDA, M. LEPELLEY, S. ARTEMOVA, J.-L. BOSSON. *Scalable Machine Learning for Predicting At-Risk Profiles Upon Hospital Admission*, in "Big Data Research", March 2018, vol. 12, pp. 23-34 [*DOI :* 10.1016/J.BDR.2018.02.004], https://hal.inria.fr/hal-01517087

[7] T. MICHEL, P. GENEVÈS, H. FOURATI, N. LAYAÏDA. *Attitude Estimation for Indoor Navigation and Augmented Reality with Smartphones*, in "Pervasive and Mobile Computing", March 2018 [*DOI :* 10.1016/J.PMCJ.2018.03.004], https://hal.inria.fr/hal-01650142

[8] P. SENELLART, L. JACHIET, S. MANIU, Y. RAMUSAT. *ProvSQL: Provenance and Probability Management in PostgreSQL*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2018, vol. 11, n° 12, pp. 2034-2037 [*DOI :* 10.14778/3229863.3236253], https://hal.inria.fr/hal-01851538

International Conferences with Proceedings

[9] A. ABBAS, P. GENEVÈS, C. ROISIN, N. LAYAÏDA. *Selectivity Estimation for SPARQL Triple Patterns with Shape Expressions*, in "ICWE'18 - 18th International Conference on Web Engineering", Cáceres, Spain, Springer, June 2018, pp. 195-209 [*DOI :* 10.1007/978-3-319-91662-0_15], https://hal.inria.fr/hal-01673013

[10] J. DAVID, J. EUZENAT, P. GENEVÈS, N. LAYAÏDA. *Evaluation of Query Transformations without Data*, in "WWW 2018 - Companion of The Web Conference", Lyon, France, ACM Press, April 2018, pp. 1599-1602 [*DOI :* 10.1145/3184558.3191617], https://hal.inria.fr/hal-01891182

[11] R. DELANAUX, A. BONIFATI, M.-C. ROUSSET, R. THION. *Query-based Linked Data Anonymization*, in "The 17th International Semantic Web Conference (ISWC 2018)", Monterey, United States, October 2018, pp. 530-546, https://hal.archives-ouvertes.fr/hal-01896276

[12] A. FEJZA, P. GENEVÈS, N. LAYAÏDA, J.-L. BOSSON. *Scalable and Interpretable Predictive Models for Electronic Health Records*, in "DSAA 2018 - 5th IEEE International Conference on Data Science and Advanced Analytics", Turin, Italy, IEEE, October 2018, pp. 1-10, https://hal.inria.fr/hal-01877742

[13] D. GRAUX, L. JACHIET, P. GENEVÈS, N. LAYAÏDA. *A Multi-Criteria Experimental Ranking of Distributed SPARQL Evaluators*, in "Big Data 2018 - IEEE International Conference on Big Data", Seattle, United States, IEEE, December 2018, pp. 1-10, https://hal.inria.fr/hal-01381781

[14] L. JACHIET, N. GESBERT, P. GENEVÈS, N. LAYAÏDA. *On the Optimization of Recursive Relational Queries*, in "BDA 2018, 34ème Conférence sur la Gestion de Données - Principes, Technologies et Applications", Bucarest, Romania, October 2018, https://hal.inria.fr/hal-01673025

[15] T. MICHEL, P. GENEVÈS, N. LAYAÏDA. *A Method to Quantitatively Evaluate Geo Augmented Reality Applications*, in "ISMAR 2018 - International Symposium on Mixed and Augmented Reality", Munich, Germany, October 2018, pp. 1-6, https://hal.inria.fr/hal-01890838

[16] D. SCHWAB, A. FEJZA, L. VIAL, Y. ROBERT. *The GazePlay Project: Open and Free Eye-trackers Games and a Community for People with Multiple Disabilities*, in "ICCHP 2018 - 16th International Conference on Computers Helping People with Special Needs", Linz, Austria, LNCS, Springer, July 2018, vol. 10896, pp. 254-261 [*DOI :* 10.1007/978-3-319-94277-3_41], https://hal.archives-ouvertes.fr/hal-01804271

Scientific Books (or Scientific Book chapters)

[17] A. BONIFATI, G. FLETCHER, H. VOIGT, N. YAKOVETS. *Querying Graphs*, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, October 2018, vol. 10, n° 3, pp. 1-184, https://hal.inria.fr/hal-01974379

Research Reports

[18] D. SCHWAB, A. FEJZA, L. VIAL, Y. ROBERT. *The GazePlay Project : Overview in February 2018*, LIG lab, February 2018, https://hal.archives-ouvertes.fr/hal-01981318

Other Publications

[19] H. IM, P. GENEVÈS, N. GESBERT, N. LAYAÏDA. *Backward Type Inference for XML Queries*, September 2018, working paper or preprint, https://hal.inria.fr/hal-01497857