IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université de Grenoble Alpes

Activity Report 2019

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 Claude Bernard, Professor, HDR]
Nils Gesbert [Institut polytechnique de Grenoble, Associate Professor]
Cécile Roisin [Univ Grenoble Alpes, Professor, HDR]

PhD Students

Fateh Boulmaiz [Qualifret SAS, PhD Student]
Sarah Chlyah [Inria, PhD Student]
Amela Fejza [Univ Grenoble Alpes, PhD Student]
Muideen Lawal [Univ Grenoble Alpes, PhD Student]

Technical staff

Thomas Calmant [Inria, Engineer]

Intern and Apprentice

Luisa Werner [Inria, from Oct 2019]

Administrative Assistant

Helen Pouchot-Rouge-Blanc [Inria, Administrative Assistant]

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.

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: Application to Graph Queries

Graph databases have received a lot of attention as they are particularly useful in many applications such as social networks, life sciences and the semantic web. Various languages have emerged to query graph databases, many of which embed forms of recursion which reveal essential for navigating in graphs. The relational model has benefited from a huge body of research in the last half century and that is why many graph databases rely on techniques of relational 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. We propose μ-RA, a variation of the Relational Algebra equipped with a fixpoint operator for expressing recursive relational queries. μ-RA can notably express unions of conjunctive regular path queries. Leveraging the fact that this fixpoint operator makes recursive terms more amenable to algebraic transformations, we propose new rewrite rules. These rules make it possible to generate new query execution plans, that cannot be obtained with previous approaches. We have defined the syntax and semantics of μ-RA, together with the rewriting rules that we specifically devised to tackle the optimization of recursive queries. We have also conducted practical experiments that show that the newly generated plans can provide significant performance improvements for evaluating recursive queries over graphs.

These results will be presented at the SIGMOD 2020 conference [9].

6.2. An Algebra with a Fixpoint Operator for Distributed Data Collections

We propose an algebra with a fixpoint operator which is suitable for modeling recursive computations with distributed data collections. We show that under reasonable conditions this fixpoint can be evaluated by parallel loops with one final merge rather than by a global loop requiring network overhead after each iteration. We also propose rewrite rules, showing when and how filters can be pushed through recursive terms, and how to filter inside a fixpoint before a join. Experiments with the Spark platform illustrate performance gains brought by these systematic optimizations [10].

6.3. 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 only describe the subtree structure of the given node. In 2015, Giuseppe Castagna and our team independently proposed 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. Recently, 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.

6.4. 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 are developping distributed models that are scalable and interpretable. Key insights include analyzing 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.

Preliminary results were presented at the 2018 International Conference on Data Science and Applications, and extended results have been submitted for publication consideration.

6.5. What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse.

In this work, our objective is to identify common temporal evolution profiles in biological data and to propose a semi-automated method to these patterns in a clinical data warehouse (CDW). We leveraged the CDW of the European Hospital Georges Pompidou and tracked the evolution of 192 biological parameters over a period of 17 years (for 445,0 0 0 + patients, and 131 million laboratory test results). We have identified three common profiles of evolution: discretization, breakpoints, and trends. We developed computational and statistical methods to identify these profiles in the CDW. Overall, of the 192 observed biological parameters (87,814,136 values), 135 presented at least one evolution. We identified breakpoints in 30 distinct parameters, discretizations in 32, and trends in 79. As a conclusion, we can say that our method allows the identification of several temporal events in the data. Considering the distribution over time of these events, we identified probable causes for the observed profiles: instruments or software upgrades and changes in computation formulas. We evaluated the potential impact for data reuse. Finally, we formulated recommendations to enable safe use and sharing of biological data collection to limit the impact of data evolution in retrospective and federated studies (e.g. the annotation of laboratory parameters presenting breakpoints or trends) [4].

6.6. Interactive Mapping Specification with Exemplar Tuples

While schema mapping specification is a cumbersome task for data curation specialists, it becomes unfeasible for non-expert users, who are unacquainted with the semantics and languages of the involved transformations.

In this work, we propose an interactive framework for schema mapping specification suited for non-expert users. The underlying key intuition is to leverage a few exemplar tuples to infer the underlying mappings and iterate the inference process via simple user interactions under the form of Boolean queries on the validity of the initial exemplar tuples. The approaches available so far are mainly assuming pairs of complete universal data examples, which can be solely provided by data curation experts, or are limited to poorly expressive mappings.

We present a quasi-lattice-based exploration of the space of all possible mappings that satisfy arbitrary user exemplar tuples. Along the exploration, we challenge the user to retain the mappings that fit the user's requirements at best and to dynamically prune the exploration space, thus reducing the number of user interactions. We prove that after the refinement process, the obtained mappings are correct and complete. We present an extensive experimental analysis devoted to measure the feasibility of our interactive mapping strategies and the inherent quality of the obtained mappings [2].

6.7. Schema Validation and Evolution for Graph Databases

Despite the maturity of commercial graph databases, little consensus has been reached so far on the standardization of data definition languages (DDLs) for property graphs (PG). Discussion on the characteristics of PG schemas is ongoing in many standardization and community groups. Although some basic aspects of a schema are already present in most commercial graph databases, full support is missing allowing to constraint property graphs with more or less flexibility. In this work, we show how schema validation can be enforced through homomorphisms between PG schemas and PG instances by leveraging a concise schema DDL inspired by Cypher syntax. We also briefly discuss PG schema evolution that relies on graph rewriting operations allowing to consider both prescriptive and descriptive schemas [6].

6.8. MapRepair: Mapping and Repairing under Policy Views

Mapping design is overwhelming for end users, who have to check at par the correctness of the mappings and the possible information disclosure over the exported source instance. In our tool MapRepair, we focus on the latter problem by proposing a novel practical solution to ensure that a mapping faithfully complies with a set of privacy restrictions specified as source policy views. We showcase MapRepair, that guides the user through the tasks of visualizing the results of the data exchange process with and without the privacy restrictions. MapRepair leverages formal privacy guarantees and is inherently data-independent, i.e. if a set of criteria are satisfied by the mapping statement, then it guarantees that both the mapping and the underlying instances do not leak sensitive information. Furthermore, MapRepair also allows to automatically repair an input mapping w.r.t. a set of policy views in case of information leakage. We build on various demonstration scenarios, including synthetic and real-world instances and mappings [5].

6.9. Approximate Querying on Property Graphs

Property graphs are becoming widespread when modeling data with complex structural characteristics and enhancing edges and nodes with a list of properties. We worked on the approximate evaluation of counting queries involving recursive paths on property graphs. As such queries are already difficult to evaluate over pure RDF graphs, they require an ad-hoc graph summary for their approximate evaluation on property graphs. We prove the intractability of the optimal graph summarization problem, under our algorithm's conditions. We design and implement a novel property graph summary suitable for the above queries, along with an approximate query evaluation module. Finally, we show the compactness of the obtained summaries as well as the accuracy of answering counting recursive queries on them [8].

6.10. RDF Graph Anonymization Robust to Data Linkage

Privacy is a major concern when publishing new datasets in the context of Linked Open Data (LOD). A new dataset published in the LOD is indeed exposed to privacy breaches due to the linkage to objects already present in the other datasets of the LOD. In this work, we focus on the problem of building safe anonymizations of an RDF graph to guarantee that linking the anonymized graph with any external RDF graph will not cause privacy breaches. Given a set of privacy queries as input, we study the data-independent safety problem and the sequence of anonymization operations necessary to enforce it. We provide sufficient conditions under which an anonymization instance is safe given a set of privacy queries. Additionally, we show that our algorithms for RDF data anonymization are robust in the presence of sameAs links that can be explicit or inferred by additional knowledge.

6.11. Navigating the Maze of Wikidata Query Logs

We propose an in-depth and diversified analysis of the Wikidata query logs, recently made publicly available. Although the usage of Wikidata queries has been the object of recent studies, our analysis of the query traffic reveals interesting and unforeseen findings concerning the usage, types of recursion, and the shape classification of complex recursive queries. Wikidata specific features combined with recursion let us identify a significant subset of the entire corpus that can be used by the community for further assessment. We consider and analyze the queries across many different dimensions, such as the robotic and organic queries, the presence/absence of constants along with the correctly executed and timed out queries. A further investigation that we pursue is to find, given a query, a number of queries structurally similar to the given query. We provide a thorough characterization of the queries in terms of their expressive power, their topological structure and shape, along with a deeper understanding of the usage of recursion in these logs. We make the code for the analysis available as open source [7].

6.12. Graph Generators: State of the Art and Open Challenges

The abundance of interconnected data has fueled the design and implementation of graph generators reproducing real-world linking properties, or gauging the effectiveness of graph algorithms, techniques and applications manipulating these data. We consider graph generation across multiple subfields, such as Semantic Web, graph databases, social networks, and community detection, along with general graphs. Despite the disparate requirements of modern graph generators throughout these communities, we analyze them under a common umbrella, reaching out the functionalities, the practical usage, and their supported operations. We argue that this classification is serving the need of providing scientists, researchers and practitioners with the right data generator at hand for their work. This survey provides a comprehensive overview of the state-of-the-art graph generators by focusing on those that are pertinent and suitable for several data-intensive tasks. Finally, we discuss open challenges and missing requirements of current graph generators along with their future extensions to new emerging fields [3].

6.13. A trichotomy for regular simple path queries on graphs

We focus on the computational complexity of regular simple path queries (RSPQs). We consider the following problem RSPQ(L) for a regular language L: given an edge-labeled digraph Gand two nodes xand y, is there a simple path from x to y that forms a word belonging to L? We fully characterize the frontier between tractability and intractability for RSPQ(L). More precisely, we prove RSPQ(L)is either AC0, NL-complete or NP-complete depending on the language L. We also provide a simple characterization of the tractable fragment in terms of regular expressions. Finally, we also discuss the complexity of deciding whether a language L belongs to the fragment above. We consider several alternative representations of L: DFAs, NFAs or regular expressions, and prove that this problem is NL-complete for the first representation and PSpace-complete for the other two [1].

7. Partnerships and Cooperations

7.1. Regional Initiatives

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.

8. Dissemination

8.1. Promoting Scientific Activities

8.1.1. Scientific Events: Organisation

8.1.1.1. General Chair, Scientific Chair

- Angela Bonifati is program chair of EDBT 2020, and co-chair of the Workshops of SIGMOD 2019.

8.1.1.2. 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), and co-chair and organizer of the EDBT 2019 summer school.

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 IJCAI'19, AAAI'20, IJCAI-PRICAI'20.
- A. Bonifati has been program committee member of VLDB 2019, PODS 2019, AAAI 2019, ICDE 2019, EDBT 2019, SIGMOD 2019, DEBS 2019, ICDT 2020.

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.

8.1.4. Scientific Expertise

- P. Genevès has been a scientific expert at ANRT for the CIFRE funding process.

8.1.5. Research Administration

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

- C. Roisin has been president of a Committee of Selection for an assistant position at university Grenoble-Alpes.
- 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-Rhône-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
- Master : N. Gesbert, 'Principes des systèmes de gestion de bases de données', 42 h eq TD, M1, 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
- Master : N. Gesbert, 'Introduction to lambda-calculus', 5 h eq TD, M2, UGA-Grenoble INP (MOSIG)
- 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 M2-level course 'Semantic Web: from XML to OWL' of the MOSIG program at UGA (36h)
- P. Genevès is responsible and teacher in 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 in progress: Muideen Lawal, Cost models for optimizing compilers based on mu-terms, PhD started in October 2017, co-supervised by Pierre Genevès and Nabil Layaïda.
- 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.

9. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] G. BAGAN, A. BONIFATI, B. GROZ. *A trichotomy for regular simple path queries on graphs*, in "Journal of Computer and System Sciences", December 2019, vol. 108, pp. 29-48, forthcoming [*DOI :* 10.1016/J.JCSS.2019.08.006], https://hal.inria.fr/hal-02435355

[2] A. BONIFATI, U. COMIGNANI, E. COQUERY, R. THION. *Interactive Mapping Specification with Exemplar Tuples*, in "ACM Transactions on Database Systems", June 2019, vol. 44, n^{o} 3, 44 p. , https://hal.inria.fr/hal-02096764

[3] A. BONIFATI, I. HOLUBOVÀ, A. PRAT-PÉREZ. *Graph Generators: State of the Art and Open Challenges*, in "ACM Computing Surveys", December 2019, forthcoming, https://hal.inria.fr/hal-02435371

[4] V. LOOTEN, L. KONG WIN CHANG, A. NEURAZ, M.-A. LANDAU-LORIOT, B. VEDIE, J.-L. PAUL, L. MAUGE, N. RIVET, A. BONIFATI, G. CHATELLIER, A. BURGUN, B. RANCE. *What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse*, in "Computer Methods and Programs in Biomedicine", November 2019, vol. 181, pp. 1-20, https://github.com/equipe22 [*DOI :* 10.1016/J.CMPB.2018.12.030], https://hal.archives-ouvertes.fr/hal-01978796

International Conferences with Proceedings

[5] A. BONIFATI, U. COMIGNANI, E. TSAMOURA. *Repairing mappings under policy views*, in "SIGMOD 2019 - ACM SIGMOD/PODS International Conference on Management of Data", Amsterdam, Netherlands, ACM, June 2019, pp. 1873-1876, Demonstration, https://hal.inria.fr/hal-02096750

[6] A. BONIFATI, P. FURNISS, A. GREEN, R. HARMER, E. OSHURKO, H. VOIGT. *Schema validation and evolution for graph databases*, in "ER 2019 - 38th International Conference on Conceptual Modeling", Salvador, Brazil, Springer, November 2019, pp. 448-456 [*DOI :* 10.1007/978-3-030-33223-5_37], https://hal.archives-ouvertes.fr/hal-02138771

[7] A. BONIFATI, W. MARTENS, T. TIMM. *Navigating the Maze of Wikidata Query Logs*, in "WWW 2019 - The World Wide Web Conference", San Francisco, United States, ACM, May 2019, pp. 127-138 [*DOI :* 10.1145/3308558.3313472], https://hal.inria.fr/hal-02096714

[8] S. DUMBRAVA, A. BONIFATI, A. NAZABAL RUIZ DIAZ, R. VUILLEMOT. *Approximate Querying on Property Graphs*, in "SUM 2019 - 13th international conference on Scalable Uncertainty Management", Compiègne, France, December 2019, pp. 250-265 [*DOI :* 10.1007/978-3-030-35514-2_19], https://hal.archives-ouvertes.fr/hal-02417259

[9] L. JACHIET, P. GENEVÈS, N. GESBERT, N. LAYAÏDA. *On the Optimization of Recursive Relational Queries: Application to Graph Queries*, in "SIGMOD 2020 - ACM International Conference on Management of Data", Portland, United States, June 2020, pp. 1-23, https://hal.inria.fr/hal-01673025

Other Publications

[10] S. Chlyah, N. Gesbert, P. Genevès, N. Layaïda. *An Algebra with a Fixpoint Operator for Distributed Data Collections*, March 2019, working paper or preprint, https://hal.inria.fr/hal-02066649