RESEARCH CENTRE

Inria Centre at Université Grenoble Alpes

IN PARTNERSHIP WITH:

CNRS, Université de Grenoble Alpes

2023

ACTIVITY REPORT

Project-Team TYREX

Types and Reasoning for the Web

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

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Inria

Contents

Project-Team TYREX

Creation of the Project-Team: 2014 July 01

Keywords

Computer sciences and digital sciences

A2.1.1. – Semantics of programming languages

A2.1.4. – Functional programming

A2.1.10. – Domain-specific languages

A2.2.8. – Code generation

A2.4. – Formal method for verification, reliability, certification

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.4. – Uncertain data

A3.1.6. – Query optimization

A3.1.9. – Database

A3.1.11. – Structured data

A3.2.1. – Knowledge bases

A3.2.2. – Knowledge extraction, cleaning

A3.2.3. – Inference

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.3. – Big data analysis

A3.4. – Machine learning and statistics

A3.4.1. – Supervised learning

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

A9.10. – Hybrid approaches for AI

Other research topics and application domains

B2. – Health

B6.1. – Software industry

B6.5. – Information systems

B9.5.1. – Computer science

B9.5.6. – Data science

B9.7.2. – Open data

1 Team members, visitors, external collaborators

Research Scientists

- Pierre Genevès [Team leader, CNRS, Senior Researcher, HDR]
- Nabil Layaïda [INRIA, Senior Researcher, HDR]

Faculty Members

- Ugo Comignani [GRENOBLE INP, Associate Professor]
- Nils Gesbert [GRENOBLE INP, Associate Professor]

Post-Doctoral Fellow

- Chandan Sharma [CNRS, Post-Doctoral Fellow, from Feb 2023]

PhD Students

- Guillaume Delplanque [UGA, from Oct 2023]
- Amela Fejza [UGA, ATER, until Aug 2023]
- Luisa Werner [UGA]
- Maroua Zeblah [OPENSEE SAS, CIFRE, from Apr 2023]

Technical Staff

- Sarah Chlyah [INRIA, Engineer]

Interns and Apprentices

- Guillaume Delplanque [INRIA, Intern, from Feb 2023 until Aug 2023]

Administrative Assistant

- Helen Pouchot-Rouge-Blanc [INRIA]

External Collaborators

- Laurent Carcone [W3C (ERCIM), from May 2023]
- Laurent Carcone [ERCIM, until Apr 2023]

2 Overall objectives

2.1 Objectives

We develop the foundations for the next generation of information extraction, data analysis and neuro-symbolic programming systems. Our research extends ideas from data management, artificial intelligence, programming languages and logic.

Extracting value from data increasingly requires sophisticated algorithms to represent, query, process, analyze and interpret data. We develop the foundations of data processing systems and neuro-symbolic programming, with a focus on extracting information from graph structures. These graph structures are obtained from raw data that may be more or less structured, noisy, uncertain or incomplete. Challenges include robust, efficient and scalable processing of large graphs obtained from such data. We study and

build new information extraction methods, as well as new robust and scalable programming methods for rich graph data structures.

3 Research program

3.1 Algebraic Foundations for Robust Expressive and Efficient Information Extraction

We investigate intermediate languages based on algebraic foundations for the representation, characterization, transformations and compilation of queries. We develop the algebraic and logical foundations of advanced data programming languages (extended relational algebras, algorithms, compilers) for more expressive and efficient query languages, in particular through aspects such as recursion, types, analytics, and provenance.

3.2 Neuro-Symbolic Programming

We investigate neuro-symbolic programming methods with graphs. This includes studying the integration between neural networks and symbolic logic and/or algebra. Challenges include support for rich knowledge and property graphs, and scalability issues with large practical graphs.

4 Application domains

4.1 Querying Large Graphs

Increasingly large amounts of graph-structured data become available. We develop methods which apply to the efficient evaluation of graph queries over large graphs. In particular, we consider knowledge graphs structured in the Resource Description Format (RDF) and property graphs. We develop query languages for extracting information from these graphs. We compile graph queries into the algebraic foundations that we develop, and then to lower-level code that can be executed by a variety of backends such as relational database management systems and big data frameworks such as Apache Spark. Applications of graph querying are ubiquitous: large knowledge bases, social networks, road networks, trust networks and fraud detection for cryptocurrencies, citation 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 Grenoble University Hospital (CHUGA), we explore solutions to the problem of predicting important clinical outcomes such as risks of adverse effects, nosocomial infections or inpatient mortality, based on large amounts of clinical data.

5 Social and environmental responsibility

5.1 Impact of research results

Our work on graph query optimization helps in reducing resource consumption in information extraction. Our work in neuro-symbolic programming helps in reducing the amount of data required when training accurate artificial intelligence models, thanks to the integration of a symbolic layer.

6 New software, platforms, open data

6.1 New software

6.1.1 MuIR

Name: Mu Intermediate Representation System

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. This includes query evaluators for a variety of RDBMS backends including PostgreSQL as well a distributed evaluator of algebraic terms using Apache Spark. This also includes an implementation of an efficient enumerator for recursive query plans, cost estimations, and compilers for recursive graph queries. The overall system is described in the CIKM 2023 demonstration paper.

Publications: hal-01673025, hal-03295445, hal-03004218, hal-03517826

Contact: Pierre Genevès

6.1.2 KeGNN

Name: Knowledge Enhanced Graph Neural Networks

Keywords: Artificial intelligence, Graph Neural Networks, Neural networks, Logic programming, Explainable Artificial Intelligence

Functional Description: We propose KeGNN, a neuro-symbolic framework for learning on graph data that combines both paradigms and allows for the integration of prior knowledge into a graph neural network model. In essence, KeGNN consists of a graph neural network as a base on which knowledge enhancement layers are stacked with the objective of refining predictions with respect to prior knowledge. We instantiate KeGNN in conjunction with two standard graph neural networks: Graph Convolutional Networks and Graph Attention Networks, and evaluate KeGNN on multiple benchmark datasets for node classification.

URL: https://gitlab.inria.fr/tyrex-public/kegnn

Publication: hal-04041691

Contact: Pierre Genevès

6.1.3 Reproducibility-aaai24

Keyword: Artificial intelligence

Functional Description: This is a re-implementation of the experiments conducted with Knowledge Enhanced Neural Networks (KENN) on the Citeseer Dataset, including the re-implementation of the Experiments in PyTorch and PyTorch Geometric. We also extended the experiments to the datasets Cora and PubMed.

URL: https://gitlab.inria.fr/tyrex-public/reproducibility-aaai24

Publication: hal-04035305

Contact: Pierre Genevès

6.1.4 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.

Publications: hal-01517087, hal-01877742, hal-03124966, hal-03125018, hal-03160473, hal-03066941, hal-03266004

Contact: Pierre Genevès

Partner: CHU Grenoble

7 New results

7.1 Knowledge Enhanced Graph Neural Networks for Graph Completion

Participants: Luisa Werner, Sarah Chlyah, Nabil Layaïda, Pierre Genevès.

Graph data is omnipresent and has a wide variety of applications, such as in natural science, social networks, or the semantic web. However, while being rich in information, graphs are often noisy and incomplete. As a result, graph completion tasks, such as node classification or link prediction, have gained attention. On the one hand, neural methods, such as graph neural networks, have proven to be robust tools for learning rich representations of noisy graphs. On the other hand, symbolic methods enable exact reasoning on graphs. We propose Knowledge Enhanced Graph Neural Networks (KeGNN), a neuro-symbolic framework for graph completion that combines both paradigms as it allows for the integration of prior knowledge into a graph neural network model. Essentially, KeGNN consists of a graph neural network as a base upon which knowledge enhancement layers are stacked with the goal of refining predictions with respect to prior knowledge. We instantiate KeGNN in conjunction with two state-of-the-art graph neural networks, Graph Convolutional Networks and Graph Attention Networks, and evaluate KeGNN on multiple benchmark datasets for node classification [2] [6.1.2].

7.2 Reproduce, Replicate, Reevaluate. The Long but Safe Way to Extend Machine Learning Methods

Participants: Luisa Werner, Nabil Layaïda, Pierre Genevès.

Reproducibility is a desirable property of scientific research. On the one hand, it increases confidence in results. On the other hand, reproducible results can be extended on a solid basis. In rapidly developing fields such as machine learning, the latter is particularly important to ensure the reliability of research. We present a systematic approach to reproducing (using the available implementation), replicating (using an alternative implementation) and reevaluating (using different datasets) state-of-the-art experiments. This approach enables the early detection and correction of deficiencies and thus the development of more robust and transparent machine learning methods. We detail the independent reproduction, replication, and reevaluation of initially published experiments with a method that we want to extend. For each step, we identify issues and draw lessons learned. We further discuss solutions that have proven effective in

overcoming the encountered problems. This work can serve as a guide for further reproducibility studies and generally improve reproducibility in machine learning [3] [6.1.3].

7.3 Efficient Enumeration of Recursive Plans in Transformation-based Query Optimizers

Participants: Amela Fejza, Sarah Chlyah, Nils Gesbert, Pierre Genevès, Nabil Layaïda.

Query optimizers built on the transformation-based Volcano/Cascades framework are used in many database systems. Transformations proposed earlier on the logical query dag (LQDAG) data structure, which is key in such a framework, focus only on recursion-free queries. We propose the recursive logical query dag (RLQDAG) which extends the LQDAG with the ability to capture and transform recursive queries, leveraging recent developments in recursive relational algebra. Specifically, this extension includes: (i) the ability of capturing and transforming sets of recursive relational terms thanks to (ii) annotated equivalence nodes used for guiding transformations that are more complex in the presence of recursion; and (iii) RLQDAG rewrite rules that transform sets of subterms in a grouped manner, instead of transforming individual terms in a sequential manner; and that (iv) incrementally update the necessary annotations. Core concepts of the RLQDAG are formalized using a syntax and formal semantics with a particular focus on subterm sharing and recursion. The result is a clean generalization of the LQDAG transformation-based approach, enabling more efficient explorations of plan spaces for recursive queries. An implementation of the proposed approach shows significant performance gains compared to the state-of-the-art [4, 6] [6.1.1].

7.4 The mu-RA System for Recursive Path Queries over Graphs

Participants: Amela Fejza, Sarah Chlyah, Nils Gesbert, Pierre Genevès, Nabil Layaïda.

We demonstrate a system for recursive query answering over graphs. The system is based on a complete implementation of the recursive relational algebra mu-RA, extended with parsers and compilers adapted for queries over knowledge and property graphs. Each component of the system comes with novelty for processing recursion. As a result, one can formulate, optimize and efficiently answer expressive queries that navigate recursively along paths in different types of graphs. We demonstrate the system on real datasets and show how it performs considering other state-of-the-art systems [1] [6.1.1].

7.5 Efficient Iterative Programs with Distributed Data Collections

Participants: Sarah Chlyah, Nils Gesbert, Nabil Layaïda, Pierre Genevès.

Big data programming frameworks have become increasingly important for the development of applications for which performance and scalability are critical. In those complex frameworks, optimizing code by hand is hard and time-consuming, making automated optimization particularly necessary. In order to automate optimization, a prerequisite is to find suitable abstractions to represent programs; for instance, algebras based on monads or monoids to represent distributed data collections. Currently, however, such algebras do not represent recursive programs in a way which allows for analyzing or rewriting them. In this paper, we extend a monoid algebra with a fixpoint operator for representing recursion as a first class citizen and show how it enables new optimizations. Experiments with the Spark platform illustrate performance gains brought by these systematic optimizations [5].

8 Bilateral contracts and grants with industry

8.1 Bilateral contracts with industry

Participants: Pierre Genevès, Nabil Layaïda, Maroua Zeblah.

We have a collaboration with the French Opensee fintech startup located in Paris about query optimization for multidimensional data, with a CIFRE thesis.

9 Partnerships and cooperations

9.1 National initiatives

9.1.1 ANR

GraphRec

Participants: Pierre Genevès, Nabil Layaïda, Nils Gesbert, Sarah Chlyah, Ugo Comignani, Luisa Werner, Chandan Sharma.

- Title: GraphRec: Efficient and Scalable Recursive Programming with Graphs
- ANR, Appel à projets générique 2023 – CE23 – Intelligence artificielle et science des données, PRME
- Coordinator: Pierre Genevès
- Abstract: This project seeks to design and develop novel methods for expressive and efficient information extraction from graphs, based on recursive graph queries and neuro-symbolic programming.

Newcare

Participants: Pierre Genevès, Nabil Layaïda, Luisa Werner.

- Title: Network for hEalth Workers : Covid And oRganization of Emergency teams – NEWCARE
- Duration: January 2021 – Mars 2024
- Coordinator: Marie-Estelle BINET (Laboratoire d'Economie Appliquée de Grenoble)
- Abstract: This research project has several objectives. The first one is to create an original database to describe the characteristics and interactions between caregivers working in healthcare teams in the emergency department. These data will be extracted (or desilated) from the PREDIMED clinical data warehouse (CDW), which gathers health and administrative data from patients and healthcare professionals working at Grenoble University Hospital. Then, the analysis of social networks will allow us to identify the modes of collaboration in place between caregivers and their ability to adapt to their environment. Impact evaluation methods will allow us to estimate the impact of the organizational changes caused by the covid-19 health crisis on the quality of work and the well-being of healthcare professionals.

Participation to MIAI Chairs

Participants: Pierre Genevès, Nabil Layaïda, Amela Fejza, Luisa Werner.

P. Genevès is member of the board of the DeepCare MIAI Chair. A. Fejza has participated to the DeepCare MIAI Chair. N. Layaïda, L.Werner and P. Genevès also participate to the Knowledge communication and evolution MIAI Chair.

10 Dissemination

Participants: Nils Gesbert, Ugo Comignani, Sarah Chlyah, Nabil Layaïda, Pierre Genevès.

10.1 Promoting scientific activities

10.1.1 Scientific events: selection

Member of the conference program committees Pierre Genevès has been member of the program committees of the SIGMOD 2023 and PLDI 2023 conferences.

Reviewer Nabil Layaïda has been reviewer for 2023 ACM International Conference on Information and Knowledge Management (CIKM 2023), Birmingham, UK.

10.1.2 Research administration

Pierre Genevès is co-responsible for the Computer Science Specialty at the MSTII Doctoral School of University Grenoble Alpes (ED 217).

Pierre Genevès is member of the board at Grenoble Informatics Laboratory (LIG), responsible for the research axis on formal methods, models and languages.

Nabil Layaïda is a member of the scientific committee of the LabEx PERSYVAL-lab (Pervasive Systems and Algorithms).

Nabil Layaïda is a member of the Scientific Board of Digital League, the digital cluster of Auvergne-Rhône-Alpes.

Nabil Layaïda has been president of the Hiring committe CRCN-ISFP Centre Inria de l'Université de Rennes 2023. Président du jury d'admissibilité CRCN-ISFP RBA n° 7 2023.

Nabil Layaïda has been Member of the Final Hiring committe ISFP Centre Inria de l'Université de Rennes 2023. Membre du jury d'admission ISFP RBA n° 7 2023.

Sarah Chlyah has been member of the hiring committee for the recruitement of a research engineer at Inria.

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

- Master: P. Genevès is co-responsible and teacher of the M2-level course "Fundamentals of Data Processing and Distributed Knowledge" of the MOSIG program at UGA (36h)
- Master: P. Genevès is co-responsible and teacher of the M2-level course "Accès à l'information: du web des données au web sémantique" in the ENSIMAG ISI 3A program at Grenoble-INP (30h)
- Master : N. Gesbert, Academic tutorship of an apprentice, 6 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, "Construction d'applications Web", 27 h eq TD, M1, Grenoble INP

- Master : N. Gesbert, “Principes des systèmes de gestion des bases de données”, 58 h eq TD, M1, Grenoble INP
- Master : N. Gesbert, “Introduction to lambda-calculus”, 4 h eq TD, M2, UGA-Grenoble INP (MOSIG)
- Licence : N. Gesbert, “Logique pour l’informatique”, 45 h eq TD, L3, Grenoble INP
- N. Gesbert is in charge of the L3-level course “logique pour l’informatique” and of the M1-level course “Principes des systèmes de gestion de bases de données (SEOC)”.
- Master : U. Comignani is responsible of the pedagogical team “Gestion de données” at Grenoble INP Ensimag
- Master : U. Comignani is co-responsible of the “BigData” master, co-accredited between Grenoble Ecole de Management and Grenoble INP
- Master : U. Comignani is in charge of the “Projets fil rouge”, 10 h eq TD, MS BigData, Grenoble INP
- Master : U. Comignani, “Principes des systèmes de gestion de bases de données”, 99.5 h eq TD, M1, Grenoble INP
- Master : U. Comignani is in charge of the “Projet BD”, 64 h eq TD, M1, Grenoble INP
- Master : U. Comignani, “Stockage et traitement de données à grande échelle”, 34 h eq TD, M2, Grenoble INP
- Master : U. Comignani, academic tutorship of an apprentice, 10 h eq TD, M1, Grenoble INP

10.2.2 Supervision

Pierre Genevès has been supervisor of Amela Fejza’s PhD thesis entitled “On the Optimization of Recursive Plan Enumeration with an Application to Property Graph Queries” [4].

PhD in progress: Luisa Werner, Neural Symbolic Integration for Knowledge Graphs, PhD started in October 2020, co-supervised by Nabil Layaïda and Pierre Genevès.

PhD in progress: Maroua Zeblah, Query Optimisation for column oriented databases, PhD started in April 2023, co-supervised by Pierre Genevès and Nabil Layaïda.

PhD in progress: Guillaume Delplanque, Differentiable programming for Knowledge Graphs, PhD started in September 2023, co-supervised by Pierre Genevès and Nabil Layaïda.

10.2.3 Juries

Pierre Genevès and Nabil Layaïda have been jury members of Damien Graux’s HDR entitled “Autour des données du Web Sémantique: Traitements distribués, hétérogènes et avancées”. Habilitation à diriger les recherches in Computer Science. University of Grenoble Alpes.

Nabil Layaïda has been président du jury of Adam Hegel Sánchez Ayte’s PhD thesis, Large-scale ontology-based data analytics : application to the SIDES 3.0 training platform in Medicine. PhD in computer science. University of Grenoble Alpes. 19 June 2023.

11 Scientific production

11.1 Publications of the year

International peer-reviewed conferences

[1] A. Fejza, P. Genevès, N. Layaïda and S. Chlyah. ‘The Mu-RA System for Recursive Path Queries over Graphs’. In: *32nd ACM International Conference on Information and Knowledge Management (CIKM 2023)*. 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023). Birmingham, United Kingdom, 21st Oct. 2023. DOI: 10.1145/3583780.3614756. URL: https://inria.hal.science/hal-03517826.

[2] L. Werner, N. Layaïda, P. Genevès and S. Chlyah. 'Knowledge Enhanced Graph Neural Networks for Graph Completion'. In: The 10th IEEE International Conference on Data Science and Advanced Analytics. Thessalokini, Greece, 9th Oct. 2023. URL: https://inria.hal.science/hal-04041691.

[3] L. S. Werner, N. Layaïda, P. Genevès, J. Euzenat and D. Graux. 'Reproduce, Replicate, Reevaluate.The Long but Safe Way to Extend Machine Learning Methods'. In: *Proceedings of the 38th Annual AAAI Conference on Artificial Intelligence*. AAAI 2024 - 38th Annual AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, pp. 1–9. URL: https://inria.hal.science/hal-04035305.

Doctoral dissertations and habilitation theses

[4] A. Fejza. 'On the Optimization of Recursive Plan Enumeration with an Application to Property Graph Queries'. Université Grenoble Alpes [2020-....], 11th Jan. 2023. URL: https://theses.hal.science/tel-04128256.

Reports & preprints

[5] S. Chlyah, N. Gesbert, P. Genevès and N. Layaïda. *Efficient Iterative Programs with Distributed Data Collections*. 26th May 2023. URL: https://inria.hal.science/hal-04108082.

[6] A. Fejza, P. Genevès and N. Layaïda. *Efficient Enumeration of Recursive Plans in Transformation-based Query Optimizers*. 2023. URL: https://inria.hal.science/hal-03692274.