IN PARTNERSHIP WITH:
Ecole Polytechnique

Activity Report 2019

Project-Team CEDAR

Rich Data Exploration at Cloud Scale

IN COLLABORATION WITH: Laboratoire d’informatique de l’école polytechnique (LIX)

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team CEDAR

Creation of the Team: 2016 January 01, updated into Project-Team: 2018 April 01

Keywords:

Computer Science and Digital Science:

A3.1.1. - Modeling, representation
A3.1.2. - Data management, quering and storage
A3.1.3. - Distributed data
A3.1.6. - Query optimization
A3.1.7. - Open data
A3.1.8. - Big data (production, storage, transfer)
A3.1.9. - Database
A3.2.1. - Knowledge bases
A3.2.3. - Inference
A3.2.4. - Semantic Web
A3.2.5. - Ontologies
A3.3.1. - On-line analytical processing
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4.1. - Supervised learning
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A9.1. - Knowledge
A9.2. - Machine learning

Other Research Topics and Application Domains:

B8.5.1. - Participative democracy
B9.5.6. - Data science
B9.7.2. - Open data

1. Team, Visitors, External Collaborators

Research Scientists

Ioana Manolescu [Team leader, Inria, Senior Researcher, HDR]
Oana Balalau [Inria, Starting Research Position, from Nov 2019]
Yanlei Diao [École polytechnique, Researcher]

Post-Doctoral Fellows

Mirjana Mazuran [Inria, Post-Doctoral Fellow]
Fei Song [École polytechnique, Post-Doctoral Fellow]

PhD Students

Maxime Buron [Inria, PhD Student]
Tien Duc Cao [Inria, PhD Student, until Sep 2019]
Luciano Di Palma [École polytechnique, PhD Student]
Qi Fan [China Scholarship Council, PhD Student, from Dec 2019]

Paweł Guzewicz [École polytechnique, PhD Student]
Enhui Huang [École polytechnique, PhD Student]
Vincent Jacob [École polytechnique, PhD Student, from Dec 2019]
Felix Raimundo [École polytechnique, PhD Student, until Aug 2019]
Khaled Zaouk [École polytechnique, PhD Student]

Technical staff

Felipe Cordeiro Alves Dias [Inria, Engineer, from Apr 2019 until Aug 2019]
Laurent Cetinsoy [École polytechnique, Engineer, from May 2019 until Jul 2019]
Tayeb Merabti [Inria, Engineer]
Arnab Sinha [École polytechnique, Engineer, from Apr 2019]

Interns and Apprentices

Walid Ben Naceur [École polytechnique, until Feb 2019]
Aymen Ayadi [École polytechnique, until Feb 2019]
Vincent Jacob [École polytechnique, from Apr 2019 until Oct 2019]
Irene Burger [École polytechnique, from Nov 2019]
Gauthier Guinet [École polytechnique, from Nov 2019]
Jingmao You [École polytechnique]

Administrative Assistant

Maeva Jeannot [Inria, Administrative Assistant, until Oct 2019]

Visiting Scientist

Juliana Freire [Digiteo, until Jul 2019]

External Collaborators

Julien Leblay [AIST Japan, from May 2019]
Xavier Tannier [CNRS]

2. Overall Objectives

2.1. Overall Objectives

Our research aims at **models, algorithms and tools for highly efficient, easy-to-use data and knowledge management**; throughout our research, **performance at scale** is a core concern, which we address, among other techniques, by designing algorithms for a **cloud (massively parallel)** setting. Our scientific contributions fall in three interconnected areas:

Expressive models for new applications As data and knowledge applications keep extending to novel application areas, we work to devise appropriate data and knowledge models, endowed with formal semantics, to capture such applications' needs. This work mostly concerns the domains of data journalism and journalistic fact checking;

Optimization and performance at scale This topic is at the heart of Y. Diao's ERC project "Big and Fast Data", which aims at optimization with performance guarantees for real-time data processing in the cloud. Machine learning techniques and multi-objectives optimization are leveraged to build performance models for data analytics the cloud. The same boal is shared by our work on efficient evaluation of queries in dynamic knowledge bases.

Data discovery and exploration Today's Big Data is complex; understanding and exploiting it is difficult. To help users, we explore: compact summaries of knowledge bases to abstrac their structure and help users formulate queries; interactive exploration of large relational databases; techniques for automatically discovering interesting information in knowledge bases; and keyword search techniques over Big Data sources.

3. Research Program

3.1. Scalable Heterogeneous Stores

Big Data applications increasingly involve *diverse* data sources, such as: structured or unstructured documents, data graphs, relational databases etc. and it is often impractical to load (consolidate) diverse data sources in a single repository. Instead, interesting data sources need to be exploited “as they are”, with the added value of the data being realized especially through the ability to combine (join) together data from several sources. Systems capable of exploiting diverse Big Data in this fashion are usually termed *polystores*. A current limitation of polystores is that data stays captive of its original storage system, which may limit the data exploitation performance. We work to devise highly efficient storage systems for heterogeneous data across a variety of data stores.

3.2. Semantic Query Answering

In the presence of data semantics, query evaluation techniques are insufficient as they only take into account the database, but do not provide the reasoning capabilities required in order to reflect the semantic knowledge. In contrast, (ontology-based) query answering takes into account both the data and the semantic knowledge in order to compute the full query answers, blending query evaluation and semantic reasoning.

We aim at designing efficient semantic query answering algorithms, both building on cost-based reformulation algorithms developed in the team and exploring new approaches mixing materialization and reformulation.

3.3. Multi-Model Querying

As the world’s affairs get increasingly more digital, a large and varied set of data sources becomes available: they are either structured databases, such as government-gathered data (demographics, economics, taxes, elections, ...), legal records, stock quotes for specific companies, un-structured or semi-structured, including in particular graph data, sometimes endowed with semantics (see e.g. the Linked Open Data cloud). Modern data management applications, such as data journalism, are eager to combine in innovative ways both static and dynamic information coming from structured, semi-structured, and un-structured databases and social feeds. However, current content management tools for this task are not suited for the task, in particular when they require a lenghy rigid cycle of data integration and consolidation in a warehouse. Thus, we see a need for flexible tools allowing to interconnect various kinds of data sources and to query them together.

3.4. Interactive Data Exploration at Scale

In the Big Data era we are faced with an increasing gap between the fast growth of data and the limited human ability to comprehend data. Consequently, there has been a growing demand of data management tools that can bridge this gap and help users retrieve high-value content from data more effectively. To respond to such user information needs, we aim to build interactive data exploration as a new database service, using an approach called “explore-by-example”.

3.5. Exploratory Querying of Semantic Graphs

Semantic graphs including data and knowledge are hard to apprehend for users, due to the complexity of their structure and oftentimes to their large volumes. To help tame this complexity, in prior research (2014), we have presented a full framework for RDF data warehousing, specifically designed for heterogeneous and semantic-rich graphs. However, this framework still leaves to the users the burden of chosing the most interesting warehousing queries to ask. More user-friendly data management tools are needed, which help the user discover the interesting structure and information hidden within RDF graphs. This research has benefitted from the arrival in the team of Mirjana Mazuran, as well as from the start of the PhD thesis of Paweł Guzewicz, co-advised by Yanlei Diao and Ioana Manolescu.

3.6. An Unified Framework for Optimizing Data Analytics

Data analytics in the cloud has become an integral part of enterprise businesses. Big data analytics systems, however, still lack the ability to take user performance goals and budgetary constraints for a task, collectively referred to as task objectives, and automatically configure an analytic job to achieve the objectives.

Our goal, is to come up with a data analytics optimizer that can automatically determine a cluster configuration with a suitable number of cores as well as other runtime system parameters that best meet the task objectives. To achieve this, we also need to design a multi-objective optimizer that constructs a Pareto optimal set of job configurations for task-specific objectives, and recommends new job configurations to best meet these objectives.

4. Application Domains

4.1. Cloud Computing

Cloud computing services are strongly developing and more and more companies and institutions resort to running their computations in the cloud, in order to avoid the hassle of running their own infrastructure. Today's cloud service providers guarantee machine availabilities in their Service Level Agreement (SLA), without any guarantees on performance measures according to a specific cost budget. Running analytics on big data systems require the user not to only reserve the suitable cloud instances over which the big data system will be running, but also setting many system parameters like the degree of parallelism and granularity of scheduling. Chosing values for these parameters, and chosing cloud instances need to meet user objectives regarding latency, throughput and cost measures, which is a complex task if it's done manually by the user. Hence, we need need to transform cloud service models from availabily to user performance objective rises and leads to the problem of multi-objective optimization. Research carried out in the team within the ERC project "Big and Fast Data Analytics" aims to develop a novel optimization framework for providing guarantees on the performance while controlling the cost of data processing in the cloud.

4.2. Computational Journalism

Modern journalism increasingly relies on content management technologies in order to represent, store, and query source data and media objects themselves. Writing news articles increasingly requires consulting several sources, interpreting their findings in context, and crossing links between related sources of information. CEDARresearch results directly applicable to this area provide techniques and tools for rich Web content warehouse management. Within the ANR ContentCheck project, and also as part of our international collaboration with the AIST institute from Japan, we work on one hand, to lay down foundations for computational data journalism and fact checking, and also work to devise concrete algorithms and platforms to help journalists perform their work better and/or faster. This work is carried in collaboration with Le Monde's "Les Décodeurs".

On a related topic, heterogeneous data integration under a virtual graph abstract model is studied within the ICODA Inria project which has started in September 2017. There, we collaborate with Les Décodeurs as well as with Ouest France and Agence France Presse (AFP). The data and knowledge integration framework resulting from this work will support journalists' effort to organize and analyze their knowledge and exploit it in order to produce new content.

5. Highlights of the Year

5.1. Highlights of the Year

- Through 2019 competitive **hiring**, the team has doubled its number of senior members: Oana Bălălău has been hired on an Inria Starting Researcher Position (SRP), and she joined in november; Angelos Anadiotis has been hired as a Gaspard Monge Assistant Professor at Ecole Polytechnique within the team.
- I. Manolescu and M. Buron have demonstrated the **ConnectionLens** system to the Defense Minister Florence Parly, as part of DataIA's showing for her visit at Inria, in April 2019 [1]. The national Inria director Bruno Sportisse, the military director of Ecole polytechnique François Bouchet, and the Fields medalist Cédric Villani were also present.
- As a member of the scientific committee of the **GFAIH** (Global Forum on AI for Humanity), I. Manolescu had the opportunity to meet, in a dinner at the Elysée Palace, and exchange with the French President Emmanuel Macron, the Economy and Industry Minister Bruno Le Maire, the Research Minister Frédérique Vidal, and the Digital Affairs Minister Cedric O [2].

5.1.1. Awards

- The demonstration "Spade: A Modular Framework for Analytical Exploration of RDF Graphs"[15] has obtained the **Best Demonstration Award** at the BDA conference 2019, where it has also been informally presented [3].

6. New Software and Platforms

6.1. Tatooine

KEYWORDS: RDF - JSon - Knowledge database - Databases - Data integration - Polystore

FUNCTIONAL DESCRIPTION: Tatooine allows to jointly query data sources of heterogeneous formats and data models (relations, RDF graphs, JSON documents etc.) under a single interface. It is capable of evaluating conjunctive queries over several such data sources, distributing computations between the underlying single-data model systems and a Java-based integration layer based on nested tuples.

- Participants: François Goasdoué, Ioana Manolescu, Javier Letelier Ruiz, Michaël Thomazo, Oscar Santiago Mendoza Rivera, Raphael Bonaque, Swen Ribeiro, Tien Duc Cao and Xavier Tannier
- Contact: Ioana Manolescu

6.2. AIDES

KEYWORDS: Data Exploration - Active Learning

FUNCTIONAL DESCRIPTION: AIDES is a data exploration software. It allows a user to explore a huge (tabular) dataset and discover tuples matching his or her interest. Our system repeatedly proposes the most informative tuples to the user, who must annotate them as "interesting" / "not-interesting", and as iterations progress an increasingly accurate model of the user's interest region is built. Our system also focuses on supporting low selectivity, high-dimensional interest regions.

- Contact: Yanlei Diao

6.3. OntoSQL

KEYWORDS: RDF - Semantic Web - Querying - Databases

[1] https://team.inria.fr/cedar/connectionlens/
[2] https://twitter.com/ioanamanol/status/1189478849651904513
[3] https://twitter.com/cedarinrialix/status/1185203276142256128

FUNCTIONAL DESCRIPTION: OntoSQL is a tool providing three main functionalities: - Loading RDF graphs (consisting of data triples and possibly a schema or ontology) into a relational database, - Saturating the data based on the ontology. Currently, RDF Schema ontologies are supported. - Querying the loaded data using conjunctive queries. Data can be loaded either from distinct files or from a single file containing them both. The loading process allows to choose between two storage schemas: - One triples table. - One table per role and concept. Querying provides an SQL translation for each conjunctive query according to the storage schema used in the loading process, then the SQL query is evaluated by the underlying relational database.

- Participants: Ioana Manolescu, Michaël Thomazo and Tayeb Merabti
- Partner: Université de Rennes 1
- Contact: Ioana Manolescu
- URL: https://ontosql.inria.fr/

6.4. ConnectionLens

KEYWORDS: Data management - Big data - Information extraction - Semantic Web

FUNCTIONAL DESCRIPTION: ConnectionLens treats a set of heterogeneous, independently authored data sources as a single virtual graph, whereas nodes represent fine-granularity data items (relational tuples, attributes, key-value pairs, RDF, JSON or XML nodes…) and edges correspond either to structural connections (e.g., a tuple is in a database, an attribute is in a tuple, a JSON node has a parent…) or to similarity (sameAs) links. To further enrich the content journalists work with, we also apply entity extraction which enables to detect the people, organizations etc. mentioned in text, whether full-text or text snippets found e.g. in RDF or XML. ConnectionLens is thus capable of finding and exploiting connections present across heterogeneous data sources without requiring the user to specify any join predicate.

- Contact: Manolescu Ioana
- Publication: ConnectionLens: Finding Connections Across Heterogeneous Data Sources
- URL: https://team.inria.fr/cedar/connectionlens/

6.5. INSEE-Extract

Spreadsheets extractor

KEYWORDS: RDF - Data extraction

FUNCTIONAL DESCRIPTION: Extract content of spreadsheets automatically and store it as RDF triples

- Participants: Ioana Manolescu, Xavier Tannier and Tien Duc Cao
- Contact: Tien Duc Cao
- Publication: Extracting Linked Data from statistic spreadsheets
- URL: https://gitlab.inria.fr/cedar/excel-extractor

6.6. INSEE-Search

KEYWORDS: Document ranking - RDF

FUNCTIONAL DESCRIPTION: Searching for relevant data cells (or data row/column) given a query in natural language (French)

- Participants: Ioana Manolescu, Xavier Tannier and Tien Duc Cao
- Contact: Tien Duc Cao
- Publications: Extracting Linked Data from statistic spreadsheets - Searching for Truth in a Database of Statistics

6.7. RDFQuotient

Quotient summaries of RDF graphs

KEYWORDS: RDF - Graph algorithmics - Graph visualization - Graph summaries - Semantic Web

FUNCTIONAL DESCRIPTION: RDF graphs can be large and heterogeneous, making it hard for users to get acquainted with a new graph and understand whether it may have interesting information. To help users figure it out, we have devised novel equivalence relations among RDF nodes, capable of recognizing them as equivalent (and thus, summarize them together) despite the heterogeneity often exhibited by their incoming and outgoing node properties. From these relations, we derive four novel summaries, called Weak, Strong, Typed Weak and Typed Strong, and show how to obtain from them compact and enticing visualizations.

- Participants: Ioana Manolescu, Pawel Guzewicz and François Goasdoué
- Partner: Université de Rennes 1
- Contact: Manolescu Ioana
- Publications: hal-01325900v6 - Structural Summarization of Semantic Graphs

6.8. AIDEme

KEYWORDS: Active Learning - Data Exploration

SCIENTIFIC DESCRIPTION: AIDEme is a large-scale interactive data exploration system that is cast in a principled active learning (AL) framework: in this context, we consider the data content as a large set of records in a data source, and the user is interested in some of them but not all. In the data exploration process, the system allows the user to label a record as “interesting” or “not interesting” in each iteration, so that it can construct an increasingly-more-accurate model of the user interest. Active learning techniques are employed to select a new record from the unlabeled data source in each iteration for the user to label next in order to improve the model accuracy. Upon convergence, the model is run through the entire data source to retrieve all relevant records.

A challenge in building such a system is that existing active learning techniques experience slow convergence in learning the user interest when such exploration is performed on large datasets: for example, hundreds of labeled examples are needed to learn a user interest model over 6 attributes, as we showed using a digital sky survey of 1.9 million records. AIDEme employs a set of novel techniques to overcome the slow convergence problem:

• Factorization: We observe that a user labels a data record, her decision making process often can be broken into a set of smaller questions, and the answers to these questions can be combined to derive the final answer. This insight, formally modeled as a factorization structure, allows us to design new active learning algorithms, e.g., factorized version space algorithms [2], that break the learning problem into subproblems in a set of subspaces and perform active learning in each subspace, thereby significantly expediting convergence.

• Optimization based on class distribution: Another interesting observation is that when projecting the data space for exploration onto a subset of dimensions, the user interest pattern projected onto such a subspace often entails a convex object. When such a subspatial convex property holds, we introduce a new “dual-space model” (DSM) that builds not only a classification model from labeled examples, but also a polytope model of the data space that offers a more direct description of the areas known to be positive, areas known to be negative, and areas with unknown labels. We use both the classification model and the polytope model to predict unlabeled examples and choose the best example to label next. • Formal results on convergence: We further provide theoretical results on the convergence of our proposed techniques. Some of them can be used to detect convergence and terminate the exploration process. • Scaling to large datasets: In many applications the dataset may be too large to fit in memory. In this case, we introduce subsampling procedures and provide provable results that guarantee the performance of the model learned from the sample over the entire data source.

FUNCTIONAL DESCRIPTION: There is an increasing gap between fast growth of data and limited human ability to comprehend data. Consequently, there has been a growing demand for analytics tools that can bridge this gap and help the user retrieve high-value content from data. We introduce AIDEme, a scalable interactive data exploration system for efficiently learning a user interest pattern over a large dataset. The system is cast in a principled active learning (AL) framework, which iteratively presents strategically selected records for user labeling, thereby building an increasingly-more-accurate model of the user interest. However, a challenge in building such a system is that existing active learning techniques experience slow convergence when learning the user interest on large datasets. To overcome the problem, AIDEme explores properties of the user labeling process and the class distribution of observed data to design new active learning algorithms, which come with provable results on model accuracy, convergence, and approximation, and have evaluation results showing much improved convergence over existing AL methods while maintaining interactive speed.

RELEASE FUNCTIONAL DESCRIPTION: Project code can be found over: https://gitlab.inria.fr/ldipalma/aideme

- Participants: Luciano Di Palma and Enhui Huang
- Contact: Yanlei Diao
- URL: http://www.lix.polytechnique.fr/aideme

7. New Results

7.1. Quotient summaries of RDF graphs

We have continued and finalized our work on the question of efficiently computing informative summaries of large, heterogeneous RDF graphs. Such summaries simplify the users' efforts to understand and grasp the content of an RDF graph with which they are not familiar. For instance, Figure 1 shows the summary constructed fully automatically out of a benchmark graph of a bit more than 100 million triples.

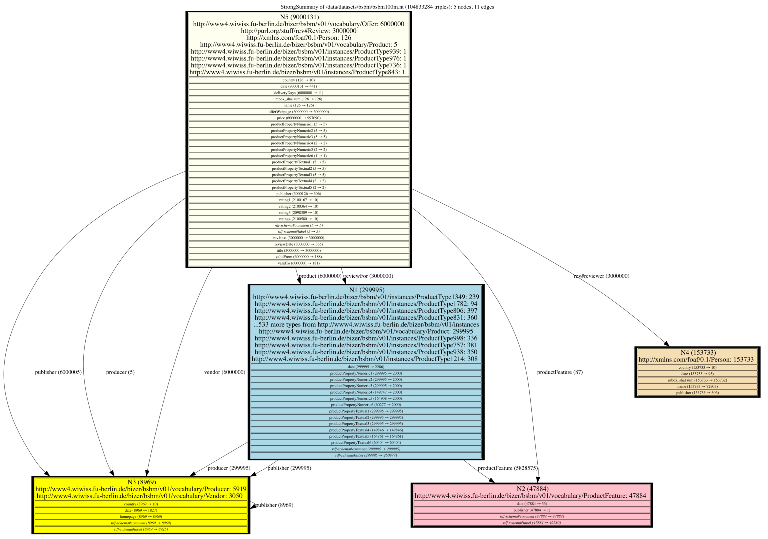

StrongSummary of /data/datasets/bsbm/bsbm100m.nt (104833284 triples): 5 nodes, 11 edges
N5 (9000131)
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/vocabulary/Offer: 6000000
http://purl.org/stuff/rev#Review: 3000000
http://xmlns.com/foaf/0.1/Person: 126
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/vocabulary/Product: 5
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType939: 1
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType976: 1
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType736: 1
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType843: 1
country (126 → 10)
date (9000131 → 441)
deliveryDays (6000000 → 11)
mbox_sha1sum (126 → 126)
name (126 → 126)
offerWebpage (6000000 → 6000000)
price (6000000 → 997090)
productPropertyNumeric1 (5 → 5)
productPropertyNumeric2 (5 → 5)
productPropertyNumeric3 (5 → 5)
productPropertyNumeric4 (2 → 2)
productPropertyNumeric5 (2 → 2)
productPropertyNumeric6 (1 → 1)
productPropertyTextual1 (5 → 5)
productPropertyTextual2 (5 → 5)
productPropertyTextual3 (5 → 5)
productPropertyTextual4 (2 → 2)
productPropertyTextual5 (2 → 2)
publisher (3000126 → 306)
rating1 (2100167 → 10)
rating2 (2100364 → 10)
rating3 (2098309 → 10)
rating4 (2100580 → 10)
rdf-schema#comment (5 → 5)
rdf-schema#label (5 → 5)
rev#text (3000000 → 3000000)
reviewDate (3000000 → 365)
title (3000000 → 3000000)
validFrom (6000000 → 188)
validTo (6000000 → 181)
product (6000000)
reviewFor (3000000)
rev#reviewer (3000000)
N1 (299995)
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType1349: 239
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType1782: 94
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType806: 397
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType831: 360
...533 more types from http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/vocabulary/Product: 299995
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType998: 336
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType757: 381
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType938: 350
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/instances/ProductType1214: 308
date (299995 → 2286)
productPropertyNumeric1 (299995 → 2000)
productPropertyNumeric2 (299995 → 2000)
productPropertyNumeric3 (299995 → 2000)
productPropertyNumeric4 (149747 → 2000)
productPropertyNumeric5 (164908 → 2000)
productPropertyNumeric6 (60277 → 2000)
productPropertyTextual1 (299995 → 299995)
productPropertyTextual2 (299995 → 299995)
productPropertyTextual3 (299995 → 299995)
productPropertyTextual4 (149846 → 149846)
productPropertyTextual5 (164861 → 164861)
productPropertyTextual6 (60404 → 60404)
rdf-schema#comment (299995 → 299995)
rdf-schema#label (299995 → 260477)
publisher (6000005)
producer (5)
vendor (6000000)
productFeature (87)
N4 (153733)
http://xmlns.com/foaf/0.1/Person: 153733
country (153733 → 10)
date (153733 → 95)
mbox_sha1sum (153733 → 153732)
name (153733 → 72903)
publisher (153733 → 306)
producer (299995)
publisher (299995)
productFeature (5828575)
N3 (8969)
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/vocabulary/Producer: 5919
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/vocabulary/Vendor: 3050
country (8969 → 10)
date (8969 → 1827)
homepage (8969 → 8969)
rdf-schema#comment (8969 → 8969)
rdf-schema#label (8969 → 8925)
publisher (8969)
N2 (47884)
http://www4.wiwiss.fu-berlin.de/bizer/bsbm/v01/vocabulary/ProductFeature: 47884
date (47884 → 33)
publisher (47884 → 1)
rdf-schema#comment (47884 → 47884)
rdf-schema#label (47884 → 46530)

Figure 1. RDFQuotient summary of a 100 million triples graph.

We have presented, together with co-authors, a tutorial on the problem of summarizing RDF graphs, at the EDBT 2019 conference [21].

We have demonstrated new algorithms for efficiently building RDF quotient summaries out of large RDF graphs, in an incremental fashion, in [19].

Last but not least, a VLDB Journal submitted article systematizing most of our contributions in this area has been accepted (pending a minor, strictly cosmetic revision which will be sent out in January 2020).

7.2. Efficient query answering over semantic graphs

Query answering in RDF knowledge bases has traditionally been performed either through graph saturation, that is, adding all implicit triples to the graph, or through query reformulation, i.e. modifying the query to look for the explicit triples entailing precisely what the original query asks for. The most expressive fragment of RDF for which reformulation-based quey answering exists is the so-called database fragment of RDF (Goasdoué et al., EDBT 2013), in which implicit triples are restricted to those entailed using an RDFS ontology. Within this fragment, query answering was so far limited to the interrogation of data triples (non-RDFS ones); however, a powerful feature specific to RDF is the ability to query data and schema triples together. In [12], we address the general query answering problem by reducing it, through a pre-query reformulation step, to that solved by the query reformulation technique mentioned above (EDBR 2013). Our experiments also demonstrate the very modest cost (performance overhead) of this more powerful (more expressive) reformulation algorithm.

7.3. Scalable storage for polystores

Big data applications routinely involve diverse datasets: relations flat or nested, complex-structure graphs, documents, poorly structured logs, or even text data. To handle the data, application designers usually rely on several data stores used side-by-side, each capable of handling one or a few data models (e.g., many relational stores can also handle JSON data), and each very efficient for some, but not all, kinds of processing on the data.

A current limitation is that applications are written taking into account which part of the data is stored in which store and how. This fails to take advantage of (i) possible redundancy, when the same data may be accessible (with different performance) from distinct data stores; (ii) partial query results (in the style of materialized views) which may be available in the stores. If data migrates to another store, to take advantage of its performance for a specific task, applications must be re-written; this is tedious and error-prone.

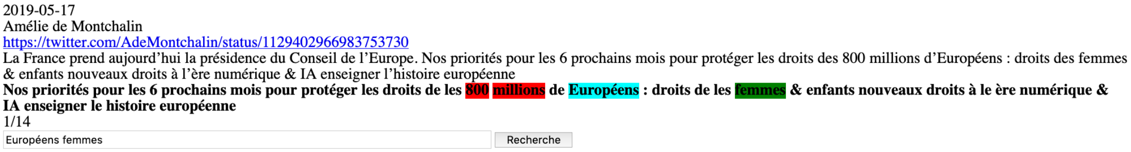
In [11], we present ESTOCADA, a novel approach connecting applications to the potentially heterogeneous systems where their input data resides. ESTOCADA can be used in a polystore setting to transparently enable each query to benefit from the best combination of stored data and available processing capabilities. ESTOCADA leverages recent advances in the area of view-based query rewriting under constraints, which we use to describe the various data models and stored data. Our experiments illustrate the significant performance gains achieved by ESTOCADA.

7.4. Novel fact-checking architectures and algorithms

A frequent journalistic fact-checking scenario is concerned with the **analysis of statements** made by individuals, whether in public or in private contexts, and the propagation of information and hearsay ("who said/knew what when"), mostly in the public sphere (e.g., in discourses, statements to the media, or on public social networks such as Twitter), but also in private contexts (these become accessible to journalists through their sources). Inspired by our collaboration with fact-checking journalists from Le Monde, France's leading newspaper, we have described in [17] a Linked Data (RDF) model, endowed with formal foundations and semantics, for describing *facts, statements*, and *beliefs*. Our model combines temporal and belief dimensions to trace propagation of knowledge between agents along time, and can answer a large variety of interesting questions through RDF query evaluation. A preliminary feasibility study of our model incarnated in a corpus of tweets demonstrates its practical interest.

Based on the above model, we implemented and demonstrated BeLink [13], a prototype capable of storing such interconnected corpora, and answer powerful queries over them relying on SPARQL 1.1. The demo showcased the exploration of a rich real-data corpus built from Twitter and mainstream media, and interconnected through extraction of statements with their sources, time, and topics.

Statistic (numerical) data, e.g., on unemployment rates or immigrant populations, are hot fact-checking topics. In prior work, we have transformed a corpus of high-quality statistics from INSEE, the French national statistics institute, into an RDF dataset (Cao et al., Semantic Big Data Workshop, 2017, https://hal.inria.fr/hal-01583975), and shown how to locate inside the information most relevant to (thus, most likely to be useful to fact-check) a given keyword query (Cao et al., Web and Databases Workshop, 2018, https://hal.inria.fr/hal-01745768). Following on the above work, in [16], we present a novel approach to extract from text documents, e.g., online media articles, mentions of statistic entities from a reference source. A claim states that an entity has certain value, at a certain time. This completes a fact-checking pipeline from text, to the reference data closest to the claim. Using it, fact-checking journalists only have to interpret the difference between the claimed and the reference value. We evaluated our method on the INSEE reference dataset and show that it is efficient and effective. Further, this algorithm was adapted also to the (more challenging) context of content published on Twitter. This has lead to a semi-automatic interface for detecting statistic claims made in tweets and starting a semi-automatic fact-check of those claims, based on INSEE data. Figure 2 depicts the interface of this Twitter fact-checking system, which was shared with our Le Monde journalist partners.



2019-05-17
Amélie de Montchalin
https://twitter.com/AdeMontchalin/status/1129402966983753730
La France prend aujourd'hui la présidence du Conseil de l'Europe. Nos priorités pour les 6 prochains mois pour protéger les droits des 800 millions d'Européens : droits des femmes & enfants nouveaux droits à l'ère numérique & IA enseigner l'histoire européenne
Nos priorités pour les 6 prochains mois pour protéger les droits de les 800 millions de Européens : droits de les femmes & enfants nouveaux droits à le ère numérique & IA enseigner le histoire européenne
1/14
Européens femmes

Figure 2. Screen capture of our Twitter fact-checking module.

7.5. Semantic graph exploration through interesting aggregates

RDF graphs can be large and complex; finding out interesting information within them is challenging. One easy method for users to discover such graphs is to be shown *interesting aggregates* (under the form of two-dimensional graphs, i.e., bar charts), where interestingness is evaluated through statistics criteria. While well understood for relational data, such exploration raises multiple challenges for RDF: facts, dimensions and measures have to be *identified* (as opposed to known beforehand); as there are more candidate aggregates, assessing their interestingness can be very costly; finally, *ontologies* bring novel specific challenges through the presence of *implicit* data, but also novel opportunities, enabling *ontology-driven exploration* from an aggregate initially proposed by the system.

The system Dagger we had previously proposed (2017) pioneered this approach, however its is quite inefficient, in particular due to the need to evaluate numerous, expensive aggregation queries.

In 2019, we have built upon Dagger to develop more efficient and more expressive versions thereof. Thus:

- In [22], we describe Dagger$^+$, which builds upon Dagger and leverages *sampling* to speed up the evaluation of potentially interesting agregates. We show that Dagger$^+$ achieves very significant execution time reductions, while reaching results very close to those of the original, less efficient system.
- Going beyond the expressive power of (candidate aggregates enumerated by) Dagger, we have developed and demonstrated [15] Spade, a *generic, extensible framework*, which we instantiated with:

(i) novel methods for enumerating candidate measures and dimensions in the vast space of possibilities provided by an RDF graph; (ii) a set of aggregate interestingness functions; (iii) ontology-based interactive exploration, and (iv) efficient early-stop techniques for estimating the interestingness of an aggregate query. A multi-dimensional aggregate automatically identified by SPADE appears in Figure 3.

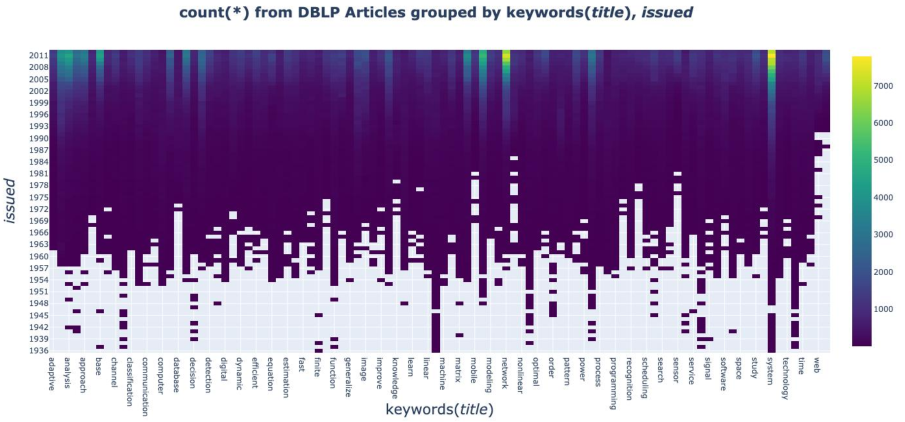


Figure 3. Interesting multi-dimensional aggregate automatically identified by DAGGER.

7.6. A Next-Generation Unified Data Analytics Optimizer

Big data analytics systems today still lack the ability to take user performance goals and budgetary constraints, collectively referred to as "objectives", and automatically configure an analytic job to achieve the objectives.

In [10], we present a unified data analytics optimizer that can automatically determine the parameters of the runtime system, collectively called a job configuration, for general dataflow programs based on user objectives. UDAO embodies key techniques including in-situ modeling, which learns a model for each user objective in the same computing environment as the job is run, and multi-objective optimization, which computes a Pareto optimal set of job configurations to reveal tradeoffs between different objectives.

Using benchmarks developed based on industry needs, our demonstration will allow the user to explore (1) learned models to gain insights into how various parameters affect user objectives; (2) Pareto frontiers to understand interesting tradeoffs between different objectives and how a configuration recommended by the optimizer explores these tradeoffs; (3) end- to-end benefits that UDAO can provide over default configurations or those manually tuned by engineers.

We demonstrated this work at the VLDB 2019 conference.

7.7. A factorized version space algorithm for interactive database exploration

One challenge in building an interactive database exploration system is that existing active learning (AL) techniques experience slow convergence when learning the user interest on large datasets. To address this slow convergence problem, we augmented version space-based AL algorithms, which have strong theoretical results on convergence but are very costly to run, with additional insights obtained in the user labeling process. These insights lead to a novel algorithm that factorizes the version space to perform active learning in a set of subspaces, with provable results on optimality, as well as optimizations for better performance. Evaluation results using real world datasets show that our algorithm significantly outperforms state-of-the-art version space algorithms, as well as our previous data exploration algorithm DSM (Huang et al., PVLDB 2018), for large database exploration.

The above work was accepted as a conference paper at ICDM 2019 [14]. In addition, we have presented a demonstration of our software at NeurIPS 2019 [26], where people could interact with our system over two real-world datasets, and also observe how our system compares against traditional AL algorithms.

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

- AIDE ("A New Database Service for Interactive Exploration on Big Data") is an ANR "Young Researcher" project led by Y. Diao, started at the end of 2016.
- ContentCheck (2015-2018) is an ANR project led by I. Manolescu, in collaboration with U. Rennes 1 (F. Goasdoué), INSA Lyon (P. Lamarre), the LIMSI lab from U. Paris Sud, and the Le Monde newspaper, in particular their fact-checking team Les Décodeurs. Its aim is to investigate content management models and tools for journalistic fact-checking.
- CQFD (2019-2022) is an ANR project coordinated by F. Ulliana (U. Montpellier), in collaboration with U. Rennes 1 (F. Goasdoué), Inria Lille (P. Bourhis), Institut Mines Télécom (A. Amarilli), Inria Paris (M. Thomazo) and CNRS (M. Bienvenu). Its research aims at investigating efficient data management methods for ontology-based access to heterogeneous databases (polystores).

8.1.2. Others

- The goal of the iCODA project is to develop the scientific and technological foundations for knowledge- mediated user-in-the-loop collaborative data analytics on heterogenous information sources, and to demonstrate the effectiveness of the approach in realistic, high-visibility use-cases. The project stands at the crossroad of multiple research fields—content analysis, data management, knowledge represen- tation, visualization—that span multiple Inria themes, and counts on a club of major press partners to define usage scenarios, provide data and demonstrate achievements. This is a project funded directly by Inria ("Inria Project Lab"), and is in collaboration with GraphIK, ILDA, LINKMEDIA (coordinator), as well as the press partners AFP, Le Monde (Les Décodeurs) and Ouest-France.

8.2. European Initiatives

8.2.1. FP7 & H2020 Projects

IDEAA: Issue-Driven European Arena Analytics is a project funded by the European Commission Union's Horizon 2020 research and innovation programme. The project started in July 2018 for a duration of two years. Its purpose is to allow citizens to easily explore the trove of publicly available data with the aim of building a viewpoint on specific issues. Its main strengths are: supply users with succinct and meaningful knowledge with respect to the issue they are interested in; allow users to interact with the provided knowledge to refine their information need and advance understanding; suggest interesting or unexpected aspects in the data and support the comparison of knowledge discovered from different data sources. IDEAA is inspired by human-to-human dialogues, where questions are explorative, possibly imprecise, and answers may be a bit inaccurate but suggestive, conveying an idea that stimulates the interlocutor to further questions.

The project supports a two-years presence of Mirjana Mazuran as an experienced post-doc in our team.

8.3. International Initiatives

8.3.1. Inria Associate Teams Not Involved in an Inria International Labs

8.3.1.1. WebClaimExplain

Title: Mining for explanations to claims published on the Web

International Partner (Institution - Laboratory - Researcher):

AIST (Japan) - Julien Leblay

Start year: 2017

See also: https://team.inria.fr/cedar/projects/webclaimexplain/

The goal of this research is to create tools to find explanations for facts and verify claims made online. While this process cannot be fully automated, the main focus of our work will be explanation finding via trusted sources, based on the observation that one can only trust a statement if he/she can explain it through rules and proofs that can themselves be trusted.

8.3.2. Inria International Partners

8.3.2.1. Informal International Partners

- We collaborate with Alin Deutsch and Rana Al-Otaibi from the University of California in San Diego, on the topic of efficient data management in polystore sytems.
- We collaborate with Helena Galhardas from the University of Lisbon on the topic of efficiently interconnecting heterogeneous data sources for journalistic applications.
- We collaborate with Anna Liu from U. Massachussets at Amherst; she co-advises PhD thesis of several students in the group (E. Huang and L. Di Palma).

8.3.3. Participation in International Programs

8.3.3.1. AYAME

WebClaimExplain

Title: Mining for explanations to claims published on the Web

International Partner (Institution - Laboratory - Researcher):

AIST (Japan) - Leblay Julien

Duration: 2017 - 2019

Start year: 2017

See also: https://team.inria.fr/cedar/connectionlens/

The goal of this research is to create tools to find explanations for facts and verify claims made online. While this process cannot be fully automated, the main focus of our work will be explanation finding via trusted sources, based on the observation that one can only trust a statement if he/she can explain it through rules and proofs that can themselves be trusted.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

We have hosted from January to July 2019 the sabbatical visit of Juliana Freire, a professor at the New York University and the president of the prestigious ACM SIGMOD scientific association.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events: Organisation

9.1.1.1. Member of the Organizing Committees

I. Manolescu was a steering committee member for the International Workshop on Misinformation, Computational Fact-Checking and Credible Web in conjunction with The Web Conference 2019.

I. Manolescu was a member of the scientific committee in charge of organizing the Global Forum on AI for Humanity (http://gfaih.org), an international conference organized under the patronage of the French government to discuss the impact and perspectives for AI research on science and the society at large. The conference featured an opening intervention by Cédric Villani and a closing speech by the French president Emmanuel Macron.

9.1.2. Scientific Events: Selection

9.1.2.1. Chair of Conference Program Committees

I. Manolescu has been a chair of the tutorial track at the ACM SIGMOD 2019 conference.

9.1.2.2. Member of the Conference Program Committees

I. Manolescu has been a member of the program committees of: the IEEE International Conference on Data Engineering (ICDE, demonstrations track) 2019, the DASFAA Conference 2019, the Extended Semantic Web Conference (ESWC) 2019, and the International Conference on Web Engineering (ICWE) 2019.

9.1.3. Journal

Y. Diao has been the Editor-in-Chief of the ACM SIGMOD Record.

9.1.3.1. Member of the Editorial Boards

Y. Diao has been an associate editor of the ACM Transactions on Database Systems (TODS).

I. Manolescu has been a member of the editorial board of the Proceedings of Very Large Databases (PVLDB) journal.

9.1.4. Invited Talks

I. Manolescu has given the following **keynote talks**:

- "Journalistic Dataspaces: Data Management for Journalism and Fact-Checking", keynote talk at the EDBT (Extending Database Technologies) Conference 2019 [28].
- "Computational fact-checking: problems, state of the art and perspectives", keynote at EGC (Extraction and Gestion de Connaissances, the French-speaking knowledge extraction and knowledge management conference) 2019 [27].

9.1.5. Leadership within the Scientific Community

- Y. Diao and I. Manolescu are members of the **PVLDB Endowment Board**, the entity in charge of organizing the publication of the prestigious PVLDB journal (A* in the CORE ranking) and of organizing the yearly PVLDB conference.
- Y. Diao has been the Chair of the ACM SIGMOD Research Highlight Award, a member of the ACM SIGMOD Executive Committee, and a member of the ACM SIGMOD Software Systems Award Committee.
- I. Manolescu is a member of the steering committee of **BDA**, the entity in charge of organizing: the yearly informal Bases de Données Avancées (BDA) conference, mostly attended by members of the French-speaking data management scientific community; and a summer school on Big Data Management, every two years.

9.1.6. Scientific Expertise

I. Manolescu has been part of the HCERES visiting committee of the Laboratoire Informatique de Grenoble (LIG) on December 2-4.

9.1.7. Research Administration

I. Manolescu has become the scientific director of **LabIA**, an initiative by the DINUM (Direction Interministerielle du Numérique) whose goal is to apply AI research and technology solutions to problems raised by the public administration, at the local or regional level. LabIA ran a selective application process which funded a dozen projects to be carried over by technology company (contractors) and four to be solved by research teams working together with the promoters (teams involved in public administration). The research projects funded by LabIA are respectively proposed by: the Cour de Cassation (the highest jurisdiction of the state), the Direction Générale de Controle de la Concurrence et de la Repression des Fraudes (DGCCRF, the national consumer watchdog agency), la SHOM (Service Hydrographique de la Marine, the seabed mapping service of the Marine) and the IGN (Institut Géographique National), in particular the team that is in charge of producing the detailed, dynamic information of the positioning of every fragment in the Earth crust.

I. Manolescu has been a member of **Inria Commission d'Evaluation** until the summer of 2019. As a consequence, she participated to the hiring committees for junior researchers (CRCN) of the Inria Lille and Inria Grenoble research centers, in May 2019; she has also participated to the final executive committee meeting that decided on the hires, in Paris, in June 2019.

I. Manolescu has been a member of a hiring committee that recruited a full-time Assistant Professor in Data Management at **Ecole Polytechnique**, and she has also headed another committee that recruited a part-time Assistant Professor in Data Science at Ecole Polytechnique.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

I. Manolescu is a part-time (50%) professor at Ecole Polytechnique, where she teaches:

- Master: I. Manolescu, "Database Management Systems", 52h, M1, École Polytechnique.
- Licence: I. Manolescu, "Giant Global Graph", 18h, L3, École Polytechnique.

She also teaches on appointment outside of Ecole Polytechnique:

- Master: I. Manolescu, "Architectures for Massive Data Management", 20h, M2, Université Paris-Saclay.

M. Buron and P. Guzewicz are Teaching Assistants at Ecole Polytechnique. Further, P. Guzewicz also taught 12h of lab in the M2 course "Architectures for Massive Data Management" mentioned above.

9.2.2. Supervision

PhD in progress: Maxime Buron: "Raisonnement efficace sur des grands graphes hétérogènes", since October 2017, François Goasdoué, Ioana Manolescu and Marie-Laure Mugnier (GraphIK Inria team in Montpellier)

PhD: Tien-Duc Cao, "Toward Automatic Fact-Checking of Statistic Claims", Université de Paris Saclay, 26/09/2019, Ioana Manolescu and Xavier Tannier (LIMICS, Université de Paris-Sorbonne).

PhD in progress: Ludivine Duroyon: "Data management models, algorithms & tools for fact-checking", since October 2017, François Goasdoué and Ioana Manolescu (Ludivine is in the Shaman team of U. Rennes 1 and IRISA, in Lannion)

PhD in progress: Paweł Guzewicz: "Expressive and efficient analytics for RDF graphs", since October 2018, Yanlei Diao and Ioana Manolescu.

PhD in progress: Qi Fan: "Multi-Objective Optimization for Data Analytics in the Cloud", since December 2019, Yanlei Diao.

PhD in progress: Enhui Huang: "Interactive Data Exploration at Scale", since October 2016, Yanlei Diao and Anna Liu (U. Massachussets at Amherst, USA).

PhD in progress: Vincent Jacob: "Explainable Anomaly Detection in High-Volume Stream Analytics", since December 2019, Yanlei Diao.

PhD in progress: Luciano di Palma, “New sampling algorithms and optimizations for interactive exploration in Big Data", since October 2017, Yanlei Diao and Anna Liu (U. Massachussets at Amherst, USA)

PhD in progress: Khaled Zaouk: “Performance Modeling and Multi-Objective Optimization for Data Analytics in the Cloud", since October 2017, Yanlei Diao.

9.2.3. Juries

- I. Manolescu has been part of the PhD committee of Adnène Belfodil, who defended his PhD thesis titled “Exceptional Model Mining for Behavioral Data Analysis” at INSA Lyon, on October 24, 2019.

9.3. Popularization

9.3.1. Articles and contents

I. Manolescu has been interviewed in the following general-audience media publications:

- “L’intelligence artificielle signe-t-elle la fin du journalisme ?”, Science et et Avenir special issue on IA, Sept 25 (dated November) 2019
- “Fake news: ces technologies qui les traquent”, Industrie et Technologies, Feb 5, 2019
- “Les algorithmes à l’assaut de la désinformation”, Science et Avenir, January 29, 2019
- “Les seniors partagent sept fois plus de «fake news» que les jeunes sur Facebook”, in Le Figaro, January 2019

9.3.2. Interventions

- Ioana Manolescu participated to a social science conference "Post-vérité et intox: où allons-nous?", organized by Fondation Maison des sciences de l’homme (FMSH) and Cité des Sciences, in February 2019 (presentation slides, présentation video)
- I. Manolescu presented her career and research at the “*Rendez-vous des jeunes mathématiciennes et informaticiennes*” (RJMI, a math and CS event organized for high-school female students) in October 2019.
- M. Buron, V. Jacob and I. Manolescu presented data management research to a group of 6 interns (13 years old, one-week long *stage de 3e*) in December 2019.

10. Bibliography

Major publications by the team in recent years

[1] R. Alotaibi, D. Bursztyn, A. Deutsch, I. Manolescu, S. Zampetakis. *Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue*, in "SIGMOD 2019 - ACM SIGMOD International Conference on Management of Data", Amsterdam, Netherlands, June 2019, https://hal.inria.fr/hal-02070827

[2] M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier. *Reformulation-based query answering for RDF graphs with RDFS ontologies*, in "ESWC 2019 - European Semantic Web Conference", Portoroz, Slovenia, March 2019, https://hal.archives-ouvertes.fr/hal-02051413

[3] D. Bursztyn, F. Goasdoué, I. Manolescu. *Teaching an RDBMS about ontological constraints*, in "Very Large Data Bases", New Delhi, India, September 2016, https://hal.inria.fr/hal-01354592

[4] S. CAZALENS, P. LAMARRE, J. LEBLAY, I. MANOLESCU, X. TANNIER. *A Content Management Perspective on Fact-Checking*, in "The Web Conference 2018 - alternate paper tracks "Journalism, Misinformation and Fact Checking"", Lyon, France, April 2018, pp. 565-574, https://hal.archives-ouvertes.fr/hal-01722666

[5] S. CEBIRIC, F. GOASDOUÉ, H. KONDYLAKIS, D. KOTZINOS, I. MANOLESCU, G. TROULLINOU, M. ZNEIKA. *Summarizing Semantic Graphs: A Survey*, in "The VLDB Journal", 2018, forthcoming, https://hal.inria.fr/hal-01925496

[6] Y. DIAO, P. GUZEWICZ, I. MANOLESCU, M. MAZURAN. *Spade: A Modular Framework for Analytical Exploration of RDF Graphs*, in "VLDB 2019 - 45th International Conference on Very Large Data Bases", Los Angeles, United States, Proceedings of the VLDB Endowment, Vol. 12, No. 12, August 2019 [*DOI :* 10.14778/3352063.3352101], https://hal.inria.fr/hal-02152844

[7] E. HUANG, L. PENG, L. D. PALMA, A. ABDELKAFI, A. LIU, Y. DIAO. *Optimization for active learning-based interactive database exploration*, in "Proceedings of the VLDB Endowment (PVLDB)", September 2018, vol. 12, n^{o} 1, pp. 71-84 [*DOI :* 10.14778/3275536.3275542], https://hal.inria.fr/hal-01969886

[8] A. ROY, Y. DIAO, U. EVANI, A. ABHYANKAR, C. HOWARTH, R. LE PRIOL, T. BLOOM. *Massively Parallel Processing of Whole Genome Sequence Data: An In-Depth Performance Study*, in "SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Dat", Chicago, Illinois, United States, SIGMOD '17 Proceedings of the 2017 ACM International Conference on Management of Data, ACM, May 2017, pp. 187-202 [*DOI :* 10.1145/3035918.3064048], https://hal.inria.fr/hal-01683398

Publications of the year

Articles in International Peer-Reviewed Journals

[9] S. CEBIRIC, F. GOASDOUÉ, H. KONDYLAKIS, D. KOTZINOS, I. MANOLESCU, G. TROULLINOU, M. ZNEIKA. *Summarizing Semantic Graphs: A Survey*, in "The VLDB Journal", June 2019, vol. 28, n^{o} 3, https://hal.inria.fr/hal-01925496

[10] K. ZAOUK, F. SONG, C. LYU, A. SINHA, Y. DIAO, P. SHENOY. *UDAO: A Next-Generation Unified Data Analytics Optimizer*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2019, vol. 12, n^{o} 12 [*DOI :* 10.14778/3352063.3352103], https://hal.inria.fr/hal-02267180

International Conferences with Proceedings

[11] R. ALOTAIBI, D. BURSZTYN, A. DEUTSCH, I. MANOLESCU, S. ZAMPETAKIS. *Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue*, in "SIGMOD 2019 - ACM SIGMOD International Conference on Management of Data", Amsterdam, Netherlands, June 2019, https://hal.inria.fr/hal-02070827

[12] M. BURON, F. GOASDOUÉ, I. MANOLESCU, M.-L. MUGNIER. *Reformulation-based query answering for RDF graphs with RDFS ontologies*, in "ESWC 2019 - European Semantic Web Conference", Portoroz, Slovenia, March 2019, https://hal.archives-ouvertes.fr/hal-02051413

[13] T.-D. CAO, L. DUROYON, F. GOASDOUÉ, I. MANOLESCU, X. TANNIER. *BeLink: Querying Networks of Facts, Statements and Beliefs*, in "ACM CIKM : 28th International Conference on Information and Knowledge Management", Beijing, China, November 2019 [*DOI :* 10.1145/3357384.3357851], https://hal.inria.fr/hal-02269134

[14] L. DI PALMA, Y. DIAO, A. LIU. *A Factorized Version Space Algorithm for "Human-In-the-Loop" Data Exploration*, in "ICDM - 19th IEEE International Conference in Data Mining", Beijing, China, November 2019, https://hal.inria.fr/hal-02274497

[15] Y. DIAO, P. GUZEWICZ, I. MANOLESCU, M. MAZURAN. *Spade: A Modular Framework for Analytical Exploration of RDF Graphs*, in "VLDB 2019 - 45th International Conference on Very Large Data Bases", Los Angeles, United States, Proceedings of the VLDB Endowment, Vol. 12, No. 12, August 2019 [*DOI :* 10.14778/3352063.3352101], https://hal.inria.fr/hal-02152844

[16] T. DUC CAO, I. MANOLESCU, X. TANNIER. *Extracting statistical mentions from textual claims to provide trusted content*, in "NLDB 2019 - 24th International Conference on Applications of Natural Language to Information Systems", Salford, United Kingdom, June 2019, https://hal.inria.fr/hal-02121389

[17] L. DUROYON, F. GOASDOUÉ, I. MANOLESCU. *A Linked Data Model for Facts, Statements and Beliefs*, in "International Workshop on Misinformation, Computational Fact-Checking and Credible Web", San Francisco, United States, WWW ’19 Companion - Proceedings of the 2019 World Wide Web Conference, May 2019 [*DOI :* 10.1145/3308560.3316737], https://hal.inria.fr/hal-02057980

[18] A. GHAZIMATIN, O. BALALAU, R. SAHA, G. WEIKUM. *PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems*, in "The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM’20)", Houston, Texas, United States, February 2020, https://hal.inria.fr/hal-02433443

[19] F. GOASDOUÉ, P. GUZEWICZ, I. MANOLESCU. *Incremental structural summarization of RDF graphs*, in "EDBT 2019 - 22nd International Conference on Extending Database Technology", Lisbon, Portugal, March 2019, https://hal.inria.fr/hal-01978784

[20] P. GUZEWICZ, I. MANOLESCU. *Parallel Quotient Summarization of RDF Graphs*, in "SBD 2019 - International Workshop on Semantic Big Data", Amsterdam, Netherlands, June 2019 [*DOI :* 10.1145/3323878.3325809], https://hal.inria.fr/hal-02106521

[21] H. KONDYLAKIS, D. KOTZINOS, I. MANOLESCU. *RDF graph summarization: principles, techniques and applications (tutorial)*, in "EDBT/ICDT 2019 - 22nd International Conference on Extending Database Technology - Joint Conference", Lisbonne, Portugal, March 2019, https://hal.inria.fr/hal-02081474

[22] I. MANOLESCU, M. MAZURAN. *Speeding up RDF aggregate discovery through sampling*, in "BigVis 2019 - 2nd International Workshop on Big Data Visual Exploration and Analytics", Lisbon, Portugal, March 2019, https://hal.inria.fr/hal-02065993

[23] L. TANCA, D. AZZALINI, F. AZZALINI, M. MAZURAN. *Tracking the Evolution of Financial Time Series Clusters*, in "DSMM 2019 - 5th Workshop on Data Science for Macro-modeling with Financial and Economic Datasets", Amsterdam, Netherlands, June 2019, https://hal.inria.fr/hal-02191810

National Conferences with Proceedings

[24] D. AZZALINI, F. AZZALINI, M. MAZURAN, L. TANCA. *Evolution of Financial Time Series Clusters (Discussion Paper)*, in "SEBD 2019 - 27th Italian Symposium on Advanced Database Systems", Castiglione della Pescaia (Grosseto), Italy, June 2019, https://hal.inria.fr/hal-02191794

Research Reports

[25] M. Buron, F. Goasdoué, I. Manolescu, M.-L. Mugnier. *Ontology-Based RDF Integration of Heterogeneous Data*, LIX, Ecole polytechnique ; Inria Saclay, August 2019, https://hal.inria.fr/hal-02266517

Other Publications

[26] E. Huang, L. D. Palma, L. Cetinsoy, Y. Diao, A. Liu. *AIDEme: An active learning based system for interactive exploration of large datasets*, December 2019, NeurIPS 2019, https://hal.inria.fr/hal-02430750

[27] I. Manolescu. *Computational fact-checking: Problems, state of the art and perspectives*, January 2019, 19e Conférence Francophone sur l'Extraction et Gestion de Connaissances (EGC), https://hal.inria.fr/hal-01995318

[28] I. Manolescu. *Journalistic Dataspaces: Data Management for Journalism and Fact-Checking (Keynote Talk)*, March 2019, EDBT/ICDT 2019 Joint Conference, https://hal.inria.fr/hal-02081430