IN PARTNERSHIP WITH:
Ecole Polytechnique

Activity Report 2018

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]
Yanlei Diao [Ecole Polytechnique, Senior Researcher, HDR]
Michael Thomazo [Inria, Researcher, until Mar 2018]

Post-Doctoral Fellows

Mirjana Mazuran [Inria, from Jul 2018]
Fei Song [Inria]

PhD Students

Maxime Buron [Inria]
Tien Duc Cao [Inria]
Sejla Cebiric [Inria, until Jul 2018]
Luciano Di Palma [Ecole polytechnique]

Pawel Guzewicz [Ecole polytechnique, from Oct 2018]
Enhui Huang [Ecole polytechnique]
Felix Raimundo [Ecole polytechnique]
Alexandre Sevin [Ecole polytechnique, until Apr 2018]
Khaled Zaouk [Ecole polytechnique]

Technical staff

Ahmed Abdelkafi [Ecole polytechnique, until Jul 2018]
Laurent Cetinsoy [Ecole polytechnique, from Oct 2018]
Tayeb Merabti [Inria]

Interns

Aymen Ayadi [Ecole polytechnique, from Oct 2018]
Walid Aymen Ben Naceur [Ecole polytechnique, from Oct 2018]
Pawel Guzewicz [Inria, from Mar 2018 until Aug 2018]
Minh Huong Le Nguyen [Inria, from Mar 2018 until Aug 2018]
Francesco Pierri [Ecole polytechnique, until Feb 2018]

Administrative Assistant

Maeva Jeannot [Inria]

Visiting Scientists

Hugo Cisneros [CNRS, from Apr 2018 until Aug 2018]
Juliana Freire [Digiteo, from Aug 2018]
Minh Huong Le Nguyen [Ecole polytechnique, until Feb 2018]

External Collaborators

Pawel Guzewicz [Université Paris Saclay, until Feb 2018]
Lars Kegel [Université de Dresden, until Aug 2018]
Minh Huong Le Nguyen [Ecole polytechnique, from Sep 2018]
Arnaud Stiegler [Ecole polytechnique, from May 2018 until Aug 2018]
Xavier Tannier [CNRS]
Stamatios Zampetakis [Orchestra Networks]

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.

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.

4.3. Open Data Intelligence

The Web is a vast source of information, to which more is added every day either in unstructured form (Web pages) or, increasingly, as partially structured sources of information, in particular as Open Data sets, which can be seen as connected graphs of data, most frequently described in the RDF data format recommended by the W3C. Further, RDF data is also the most appropriate format for representing structured information extracted automatically from Web pages, such as the DBPedia database extracted from Wikipedia or Google's InfoBoxes. Our work on this topic has taken place within the 4-year project ODIN, funded by the Department of Defense under the RAPID innovation programme.

4.4. Genomics

One particular case of area where the increase in data production is the more consequent is genomic data, indeed the amount of data produced doubles every 7 months. Thus we want to bring the expertise from the database and big data community to help both scale the existing algorithms and design new algorithms that are scalable from the ground up.

5. Highlights of the Year

5.1. Highlights of the Year

Conference Chair

Ioana Manolescu has been a general chair of the IEEE International Conference on Data Engineering (ICDE) 2018.

Keynotes

Ioana Manolescu has given invited keynote talks at the Extended Semantic Web Conference (ESWC) 2018 [25], and at the *34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications* (BDA) 2018 [24].

PVLDB paper

A paper on "Optimization for active learning-based interactive database exploration" by Enhui Huang and co-authors has been accepted at PVLDB 2018 [10].

Prix de stage de l'Ecole Polytechnique

Camille Chanial, third-year (M1) student at Ecole Polytechnique, has been awarded a Prix de Stage for his work on the ConnectionLens prototype [9].

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

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

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.

- Partner: Université de Rennes 1
- Contact: Manolescu Ioana
- Publications: Compact Summaries of Rich Heterogeneous Graphs - Structural Summarization of Semantic Graphs

7. New Results

7.1. Interactive Data Exploration at Scale

Building upon our prior work in active learning-based interactive database exploration system, we improved this system in terms of efficiency and effectiveness. First, we formally defined the class of user interest queries to which our proposed Dual Space Model (DSM) can bring significant improvement in accuracy. Second, we generalized the DSM to arbitrary queries by forcing our system to fall back to the traditional active learning-based techniques if the requested query properties are not satisfied. Third, we launched a user study to collect real-world datasets and user interest patterns for comparison experiments. The evaluation results showed that our new system outperformed the start-of-the-art active learning techniques and data exploration systems. Fourth, to show the robustness of our system, we added some label noise into the experiments. It turned out that our system maintained a good performance and significantly outperformed traditional active learning-based system. These results have appeared in the prestigious PVLDB journal [10]. In addition, we have been working on integrating DSM with version space algorithms and designing more advanced methods to deal with label noise. In the near feature, a new software based on our proposed techniques will be put into use for interactive database exploration.

7.2. A learning-based approach to optimizing large-scale data analytics

As part of my PhD thesis of K. Zaouk, we have proposed two neural network architectures to support in-situ modeling of user objectives in large-scale data analytics. Although conceptually these architectures can work with any big data system, the modeling of user-objectives on analytics run was applied on Spark Streaming. In our problem settings where only few traces are run whenever a new workload is submitted to the cloud, we have proposed new optimizations to improve the accuracy and efficiency of the auto-encoder based architecture. Thus, we have developed a prototype that included these neural network architectures and optimizations. This prototype was then used to evaluate a benchmark of stream analytics that we developed and instrumented on top of two clusters that collect Spark Streaming workloads' traces.

We analyzed the performance of the proposed techniques and demonstrated their performance benefits over state of the art performance modeling techniques based on machine learning (such as Ottertune used in tuning traditional RDBMS). Our latest results show that we outperform Ottertune in robustness and in our problem settings. These results consolidated in a paper "Boosting Big Data Analytics with Deep Learning Models and Optimization Methods" submitted for publication, alongside with other scientific results in multi-objective optimization contributed by the co-author Fei Song. Work on this topic continues.

7.3. Event stream analysis

As enterprise information systems are collecting event streams from various sources, the ability of a system to automatically detect anomalous events and further provide human readable explanations is of paramount importance. In a position paper [19], we argue for the need of a new type of data stream analytics that can address anomaly detection and explanation discovery in a single, integrated system, which not only offers increased business intelligence, but also opens up opportunities for improved solutions. In particular, we propose a two-pass approach to building such a system, highlight the challenges, and offer initial directions for solutions.

7.4. Quotient summarization of RDF graphs

We have continued our work on efficiently computing informative summaries of large, heterogeneous RDF graphs.

First, we have noticed that type information, when available, can be used to group RDF nodes in interesting, pertinent equivalence classes. However, the integration of type in our quotient summarization framework (presented in ISWC 2017) is not straightforward, since an RDF node may have zero, one, or more than one types. In [15], we have identified a sufficient, flexible condition under which we are able to propose a form of quotient summarization based on types, even if a node has multiple types, and even if they are not organized in a tree-shape classification, but instead in a directed acyclic graph (DAG).

In parallel, we have finalized a comprehensive survey of RDF graph summarization techniques which appeared in the VLDB Journal [8]. We have also completely re-developed our RDF graph summarization platform, in order to ensure correctness, to factorize common elements across all the summarization methods, and to implement new, incremental summarization algorithms [21]. This work has attracted significant visibility through an invited keynote at the ESWC conference [25], and through an ISWC "Resource" publication where our summaries are integrated in a LOD visual exploration portal developed by the ILDA team of Inria [17].

7.5. Semantic integration of heterogeneous data

A large amount of data sources are publicly available in *heterogeneous formats* such as relational, RDF and JSON. These data sources can share information about common entities, which the users may want to query as a single dataset, possibly exploiting also a set of semantic constraints which serve as a common integration perspective. We proposed a new approach to query such *integration* of datasources in a *global RDF graph* using an *RDFS ontology* and user-specified entailment rules. Previous approaches to query answering in the presence of knowledge involve either the materialization of inferred data, or reformulation of the query; both approches have well-known drawbacks. We introduce a new way of query answering as a reduction to view-based answering in [11]. This approach avoids both materialization in the data and query reformulation.

We have also developed an RDF Schema *reformulation algorithm* taking into account the reasoning on the ontology. This algorithm reduces query answering on data in the presence of an ontology, to query evaluation (solely on the data). In particular, this reformulation algorithm can be used to speed up query answering in the integration system mentioned above.

7.6. Fact-checking: a content management perspective

Throughout the year, we have worked within the ANR ContentCheck project to analyze and systematize computational fact-checking as a discipline of computer science; we have analyzed and classified existing works in this area, proposed a generic architecture for computational fact-checking, and highlighted perspectives in a Web Conference (formerly known as WWW) article [18] and two tutorials, presented respectively at the Web conference [16] and the PVLDB conference [7]. This work has also been featured in an invited keynote at the BDA 2018 conference [24].

7.7. Novel fact-checking architectures and algorithms

Still part of our work in ContentCheck, we have worked to devise new algorithms and architectures for data journalism and journalistic fact checking.

First, we have considered the problem of making it easy to check the accuracy of a statistic claim, in the statistic database published by INSEE, the leading french statistic institute. In prior work, we had shown how the INSEE data can be converted into a collection of open data adherent to the best practices of the W3C (RDF graphs). Following up on that work, we have proposed a novel algorithm which allows to search these RDF datasets by means of user-friendly keyword queries. Our algorithm returns ranked answers at the granularity of the RDF dataset (corresponding to a spreadsheet in a statistic dataset published by INSEE) or, when possible, at the granularity of individual cells, or line/column in a spreadsheet that best matches the user query [13], [12].

Second, we have devised a new architecture for keyword search in a polystore systems, where users ask a set of keywords, and receive results showing how occurrences of these keywords across the set of data sources can be connected. This allows identifying possibly unforeseen connections across heterogeneous data sources. We have implemented this architecture in the ConnectionLens prototype, which we demonstrated in VLDB [9] and also informally at BDA [14].

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.
- CBOD ("Cloud-Based Organizational Design") is a 4-year ANR started in 2014, coordinated by prof. Ahmed Bounfour from UPS. Its goal is to study and model the ways in which cloud computing impacts the behavior and operation of companies and organizations, with a particular focus on the cloud-based management of data, a crucial asset in many companies.
- ContentCheck (2015-2018) is an ANR project 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.

8.1.2. LabEx, IdEx

- CloudSelect is a three-years project started in October 2015. It is financed by the *Institut de la Société Numérique* (ISN) of the IDEX Paris-Saclay; it funds the PhD scholarship of S. Cebiric. The project is a collaboration with A. Bounfour from the economics department of Université Paris Sud. The project aims at exploring technical and business-oriented aspects of data mobility across cloud services, and from the cloud to outside the cloud.

8.1.3. Others

- ODIN is a four-year project started (2014-2018) funded by the Direction Générale de l'Armement, between the SemSoft company, IRISA Rennes and Cedar. The project focused on developing a complete framework for analytics on Web data, in particular taking into account uncertainty, based on Semantic Web technologies such as RDF.

- 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. Our WebClaimExplain collaboration has been particularly fruitful this year in terms of publications [9], [7], [18], [16], [14].

8.3.2. Inria International Partners

8.3.2.1. Informal International Partners

We resumed our collaboration with Prof. Alin Deutsch from University of California in San Diego (UCSD), during his invited stay at U. Paris Sud. We have completed a work (started in 2015-2016) on efficient view-based query rewriting in polystores, and submitted it to a major international conference.

8.3.3. Participation in Other 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/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.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

8.4.1.1. Sabbatical programme

Juliana Freire, a professor at NYU and the chair of the ACM SIGMOD chapter, has been a visitor on sabbatical in the team since September 2018.

8.4.1.2. Internships

Lars Kegel, a PhD student at the university of Dresden, has visited the team until August 2018. He has worked on characterizing and generating time series data for benchmarking time series management software.

8.4.2. Visits to International Teams

8.4.2.1. Research Stays Abroad

Yanlei Diao spent three months at U. Massachussets at Amherst, USA.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organisation

9.1.1.1. General Chair, Scientific Chair

Ioana Manolescu has been a general chair of the IEEE International Conference on Data Engineering (ICDE) 2018.

9.1.2. Scientific Events Selection

9.1.2.1. Member of the Conference Program Committees

Ioana Manolescu has been a member of the program committees of: the International Workshop on Semantic Big Data (SBD), in cojunction with ACM SIGMOD 2018; the Data Engineering Meets Semantic Web Workshop (DESWeb) in conjunction with IEEE ICDE 2018; the WebDB workshop, in conjunction with ACM SIGMOD 2018; the 34th *Conférence sur la Gestion de Données – Principes, Technologies et Applications* (BDA) 2018.

9.1.3. Journal

9.1.3.1. Member of the Editorial Boards

Yanlei Diao has been the Editor-in-Chief of the ACM SIGMOD Record, SIGMOD's quarterly newsletter.

Ioana Manolescu has been an associate editor of the PVLDB (Proceedings of VLDB) Journal 2018.

9.1.4. Invited Talks

Ioana Manolescu has given invited keynotes at the Extended Semantic Web Conference (ESWC) 2018 , and at the *34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications* (BDA) 2018.

9.1.5. Leadership within the Scientific Community

Yanlei Diao and Ioana Manolescu are both members of the PVLDB Endowment Board of Trustees.

Ioana Manolescu has been the president the ACM SIGMOD "Jim Gray" PhD Award Committee.

Ioana Manolescu is a member of the steering committee (*Comité de Pilotage*) of "Bases de Données Avancées" (BDA), the informal association organizing the database research community in France and french-speaking countries.

9.1.6. Research Administration

Y. Diao is on the advisory board of the Data Science Initiative (DSI), a joint center between the applied mathematics and computer science departments of Ecole Polytechnique.

Ioana Manolescu has been a member of Inria's Commission d'Evaluation and of the Bureau des Comités de Projets of Inria Saclay.

Team members have participated to the following hiring committees:

- Associate professor – Ecole Polytechnique, Yanlei Diao. Jury member.
- Full professor – INSA Lyon, Ioana Manolescu. Jury member.
- Associate professor – Ecole Polytechnique, Ioana Manolescu. Jury member.
- Inria researcher – Inria Nancy, Ioana Manolescu. Jury member.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- Master: Y. Diao is a Professor at Ecole Polytechnique, where she teaches "System for Big Data" in M1; she also teaches "Systems for Big Data Analytics" in M2 in the Data Science Master Program of Université Paris Saclay.
- Master: I. Manolescu, Architectures for Massive Data Management, 12h, M2, Université Paris-Saclay.
- Master: I. Manolescu, Database Management Systems, 52h, M1, École Polytechnique.
- Licence: I. Manolescu, Giant Global Graph, 18h, L3, École Polytechnique.

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 in progress: Tien Duc Cao: "Extraction et interconnexion de connaissances appliquée aux données journalistiques", since October 2016, Ioana Manolescu and Xavier Tannier (LIMSI/CNRS and Université de Paris Sud)

PhD interrupted in July 2018: Sejla Čebirić: “CloudSelect: Data Mobility Within, Across and Outside Clouds”, since September 2015, François Goasdoué Goasdoué and Ioana Manolescu. The student joined the industry.

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

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: Felix Raimundo: “Nouveaux algorithmes et optimisations pour l’analyse profonde du génome à l’échelle de la population”, since October 2017, Yanlei Diao and Avinash Abhyankar (New York Genome Center, USA)

PhD interrupted in April 2018: Alexandre Sevin: “Exploration interactive de données sur de grandes sources de données hétérogènes”, since October 2017, Yanlei Diao and Peter Haas (U. Massachussets at Amherst, USA). The student joined the industry.

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

Ioana Manolescu reported on the PhD thesis of Louis Jachiet (Inria Grenoble).

9.3. Popularization

9.3.1. Articles and contents

- In books/journals for the general public: Our research on computational fact-checking has been featured in Le Journal Toulousain in an article titled “La science se met au service de l’information” , in February 2018 (http://www.lejournaltoulousain.fr/societe/la-science-se-met-au-service-de-linformation-pour-detecter-les-fake-news-54906). Further, I. Manolescu has been interviewed as an expert for an article in Le Figaro, in March 2018 (http://www.lefigaro.fr/secteur/high-tech/2018/03/08/32001-20180308ARTFIG00341-sur-twitter-les-fake-news-voyagent-plus-vite-que-les-vraies-informations.php).

9.3.2. Interventions

- Ioana Manolescu participated to a televised debate on fake news and the media in “Le Grand Barouf Numérique”, a technology/society meet up organised by the city of Lille, in March 2018. The debate was broadcast by France 3 (https://france3-regions.francetvinfo.fr/hauts-de-france/nord-0/lille/debat-fake-news-medias-sont-ils-malades-1446365.html).
- Ioana Manolescu has participated to a debate on regulating the news media in “Les Journées de l’Economie”, a national event organized in Lyon, in November 2018 (http://www.touteconomie.org/index.php?arc=bv1&manif=594)

9.3.3. Creation of media or tools for science outreach

Our research on computational fact-checking has been featured in a documentary (http://www.universcience.tv/video-tech-check-des-scientifiques-face-aux-fake-news-19501.html) by UniverScience.TV, the organization responsible of preparing scientific short movies next to Cité des Sciences. The movie has been featured in an article in Le Monde (https://www.lemonde.fr/sciences/video/2018/01/24/comment-la-science-aide-a-reperer-les-fake-news_5246356_1650684.html).

10. Bibliography

Major publications by the team in recent years

[1] R. Bonaque, B. Cautis, F. Goasdoué, I. Manolescu. *Social, Structured and Semantic Search*, in "International Conference on Extending Database Technology", Bordeaux, France, March 2016, https://hal.inria.fr/hal-01277939

[2] 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

[3] 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

[4] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, M. Zneika. *Summarizing Semantic Graphs: A Survey*, in "The VLDB Journal", 2018, https://hal.inria.fr/hal-01925496

[5] 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

[6] 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

[7] S. Cazalens, J. Leblay, P. Lamarre, I. Manolescu, X. Tannier. *Computational Fact Checking: A Content Management Perspective*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2018, vol. 11, n^o 12, pp. 2110-2113 [*DOI :* 10.14778/3229863.3229880], https://hal.inria.fr/hal-01853067

[8] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, M. Zneika. *Summarizing Semantic Graphs: A Survey*, in "The VLDB Journal", 2018, https://hal.inria.fr/hal-01925496

[9] C. Chanial, R. Dziri, H. Galhardas, J. Leblay, M.-H. Le Nguyen, I. Manolescu. *ConnectionLens: Finding Connections Across Heterogeneous Data Sources*, in "Proceedings of the VLDB Endowment (PVLDB)", 2018, vol. 11, 4 p. [*DOI :* 10.14778/3229863.3236252], https://hal.inria.fr/hal-01841009

[10] 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

International Conferences with Proceedings

[11] M. BURON, F. GOASDOUÉ, I. MANOLESCU, M.-L. MUGNIER. *Rewriting-Based Query Answering for Semantic Data Integration Systems*, in "34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2018)", Bucarest, Romania, October 2018, https://hal.archives-ouvertes.fr/hal-01927282

[12] T.-D. CAO, I. MANOLESCU, X. TANNIER. *Extracting Linked Data from statistic spreadsheets*, in "Conférence sur la Gestion de Données – Principes, Technologies et Applications", Bucarest, Romania, October 2018, https://hal.inria.fr/hal-01915148

[13] T.-D. CAO, I. MANOLESCU, X. TANNIER. *Searching for Truth in a Database of Statistics*, in "WebDB 2018 - 21st International Workshop on the Web and Databases", Houston, United States, June 2018, pp. 1-6, https://hal.inria.fr/hal-01745768

[14] C. CHANIAL, R. DZIRI, H. GALHARDAS, J. LEBLAY, M.-H. LE NGUYEN, I. MANOLESCU. *ConnectionLens: Finding Connections Across Heterogeneous Data Sources*, in "34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications", Bucarest, Romania, October 2018, https://hal.inria.fr/hal-01968418

[15] P. GUZEWICZ, I. MANOLESCU. *Quotient RDF Summaries Based on Type Hierarchies*, in "DESWeb'2018 - Data Engineering meets the Semantic Web 2018", Paris, France, April 2018, https://hal.inria.fr/hal-01721163

[16] J. LEBLAY, I. MANOLESCU, X. TANNIER. *Computational fact-checking: Problems, state of the art, and perspectives*, in "The Web Conference", Lyon, France, The Web Conference 2018, April 2018, https://hal.inria.fr/hal-01791232

[17] E. PIETRIGA, H. GÖZÜKAN, C. APPERT, M. DESTANDAU, Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Browsing Linked Data Catalogs with LODAtlas*, in "ISWC 2018 - 17th International Semantic Web Conference", Monterey, United States, Springer, October 2018, pp. 137-153, https://hal.inria.fr/hal-01827766

Conferences without Proceedings

[18] 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

[19] F. SONG, B. ZHOU, Q. SUN, W. SUN, S. XIA, Y. DIAO. *Anomaly Detection and Explanation Discovery on Event Streams*, in "BIRTE2018", RIO, Brazil, August 2018, https://hal.inria.fr/hal-01970660

Research Reports

[20] E. HUANG, L. PENG, L. DI PALMA, A. ABDELKAFI, A. LIU, Y. DIAO. *Optimization for Active Learning-based Interactive Database Exploration*, Ecole Polytechnique ; University of Massachusetts Amherst, 2018, https://hal.inria.fr/hal-01870560

[21] Š. ČEBIRIĆ, F. GOASDOUÉ, P. GUZEWICZ, I. MANOLESCU. *Compact Summaries of Rich Heterogeneous Graphs*, Inria Saclay ; Université Rennes 1, July 2018, n^{o} RR-8920, pp. 1-40, https://hal.inria.fr/hal-01325900

Other Publications

[22] I. CZERESNIA ETINGER. *Summary-based optimization in semantic graph databases*, March 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01742495

[23] P. GUZEWICZ. *Internship report: Quotient RDF graph summarization*, Inria Saclay ; Telecom ParisTech, September 2018, https://hal.inria.fr/hal-01879898

[24] I. MANOLESCU. *Democracy Big Bang: What data management can(not) do for journalism*, October 2018, 34ème Conférence sur la Gestion de Données – Principes, Technologies et Applications, https://hal.inria.fr/hal-01968347

[25] I. MANOLESCU. *Structural Summarization of Semantic Graphs*, June 2018, Extended Semantic Web Conference (ESWC) , https://hal.inria.fr/hal-01808737