IN PARTNERSHIP WITH:
CNRS

Université Paris-Sud (Paris 11)

Activity Report 2015

Project-Team OAK

Database optimizations and architectures for complex large data

RESEARCH CENTER
Saclay - Île-de-France

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team OAK

Creation of the Team: 2012 April 01, updated into Project-Team: 2013 January 01, end of the Project-Team: 2015 December 31

Keywords:

Computer Science and Digital Science:

3.1.1. - Modeling, representation
3.1.2. - Data management, quering and storage
3.1.6. - Query optimization
3.1.7. - Open data
3.1.8. - Big data (production, storage, transfer)
3.1.9. - Database
3.2.1. - Knowledge bases
3.2.3. - Inference
3.2.4. - Semantic Web
3.2.5. - Ontologies
3.3.3. - Big data analysis
8.1. - Knowledge

Other Research Topics and Application Domains:

6.3.1. - Web
6.3.4. - Social Networks
6.4. - Internet of things
6.5. - Information systems
9.4.5. - Data science
9.5.10. - Digital humanities
9.7.2. - Open data

1. Members

Research Scientists

Ioana Manolescu [Team leader, Inria, Senior Researcher, HdR]
Michael Thomazo [Inria, Researcher, from Dec 2015]

Faculty Members

Nicole Bidoit [Univ. Paris XI, Professor]
Bogdan Cautis [Univ. Paris XI, Professor]
Benoit Groz [Univ. Paris XI, Associate Professor]

Engineers

Elham Akbari Azirani [Inria, until Jan 2015]
Oscar Santiago Mendoza Rivera [Inria, from Nov 2015]
Swen Ribeiro [Inria, from Nov 2015]

PhD Students

Raphael Bonaque [Inria]
Damian Bursztyn [Inria]

Sejla Cebiric [Inria]
Paul Lagree [Univ. Paris XI]
Aikaterini Tzompanaki [Univ. Paris XI]

Post-Doctoral Fellows

Francesca Bugiotti [Inria, until Feb 2015]
Ioana Ileana [Inria, until Jul 2015]
Soudip Roy Chowdhury [Univ. Paris XI, until Aug 2015]
Alessandro Solimando [Inria]
Stamatios Zampetakis [Inria]

Administrative Assistants

Thi Bui [Inria, from Oct 2015]
Maeva Jeannot [Inria]

2. Overall Objectives

2.1. Overall Objectives

Data is being created at unprecedented scale and speed, and processed in increasingly varied and complex fashion. OAK research aims at devising expressive models for flexible processing of complex data, in particular Web and social data; we also devise and develop strong software tools efficiently implementing such rich models.

The team has developed pointed expertise related to the processing of Web data (in particular XML, RDF, or social graph data), and in models and architecture for the massively parallel management of Web data.

3. Research Program

3.1. Scalable and Expressive Techniques for the Semantic Web

The Semantic Web vision of a world-wide interconnected database of *facts*, describing *resources* by means of *semantics*, is coming within reach as the W3C's RDF (Resource Description Format) data model is gaining traction. The W3C Linking Open Data initiative has boosted the publication and interlinkage of a large number of datasets on the semantic web resulting to the Linked Open Data Cloud. These datasets of billions of RDF triples have been created and published online. Moreover, numerous datasets and vocabularies from different application domains are published nowadays as RDF graphs in order to facilitate community annotation and interlinkage of both scientific and scholarly data of interest. RDF storage, querying, and reasoning is now supported by a host of tools whose scalability and expressive power vary widely. Unsurprisingly, some of the most scalable tools draw upon the existing models and architecture for managing structured data. However, such tools often ignore the semantic aspects that make RDF interesting. For what concerns the semantics, a delicate balance must be found between expressive power and the efficiency of the resulting data management algorithms.

- The team works on identifying tractable dialects of RDF, amenable to highly efficient query answering algorithms, taking into account both data and semantics.
- Another line of research investigates the usage of RDF data and semantics to help structure, organize, and enrich structured documents from social media. Based on such a rich model, we devised novel query answering algorithms which attempt to explore efficiently the rich social dataset in order to return the most pertinent answers to the users, from a social, structured and semantic perspective. This research is related to the DIGICOSME LabEx grant "Structured, Social and Semantic Search".
- To help users get acquainted with large and complex RDF graphs, we have started to work on an approach for RDF graph summarization: a graph summary is a smaller RDF graph, often by several orders of magnitude, which preserves the core structural information of the original graph and thus allows to reason about several important graph property on a much more manageable structure.

3.2. Massively Distributed Data Management Systems

Large and increasing data volumes have raised the need for distributed storage architectures. Among such architectures, computing in the cloud is an emerging paradigm massively adopted in many applications for the scalability, fault-tolerance and elasticity features it offers, which also allows for effortless deployment of distributed and parallel architectures. At the same time, interest in massively parallel processing has been renewed by the MapReduce model and many follow-up works, which aim at simplifying the deployment of massively parallel data management tasks in a cloud environment. For these reasons, cloud-based stores are an interesting avenue to explore for handling very large volumes of RDF data.

A recent development in this area is the start of our collaboration with social scientists from UNIV. PARIS-SUD, working on the management of innovation; we have started a collaborative research projects (ANR "Cloud-Based Organizational Design") where we perform an interdisciplinary analysis (both from a computing and from a business management perspective) on the adoption of cloud technologies within an enterprise.

3.3. Social Data Management and Crowdsourcing

The social Web blurs today the distinction between search, recommendation, and advertising (three paradigms for information access that have been so far considered mostly in separation). Our research in this area strives to find better adapted and scalable ways to answer information needs in the social Web, often by techniques at the intersection of databases, information retrieval, and data mining.

In particular, we study models and algorithms for personalized, or social-aware search in social applications. While progress has been made in this area, more remains to be done in order to address users' needs in practice, especially towards richer data models, and improving applicability and result relevance. For instance, when searching for tweets, their geographical location and recency may be as important for relevance as the textual and social aspects.

Furthermore, regarding quality of answers in response to searches, for various reasons (e.g., sparsity or tagging quality), meaningful results may often not be available. One response to this observation could be to turn to the crowd, the very users/publishers of the social media platform, and to turn this crowd into on-demand and query-driven sources of data. We study principled approaches for crowd selection (expert sourcing) and task assignment (data sourcing), in order to better answer ongoing social queries.

Beyond social links that represent just ties, a promising direction we also focus on in user-centric applications is to uncover implicit, potentially richer relationships from user interactions and to exploit them to improve core functionality such as search.

Moreover, we plan to investigate how crowdsourcing can be exploited to extract informations on user preferences, using techniques about noisy data management and provenance analysis.

4. Application Domains

4.1. Social Networks

We develop models and algorithms for efficiently exploiting, enhancing, and querying social network data, in particular based on structured content, semantic annotations, and user interaction networks. We pursue this research with many industrial partners within the ALICIA project (Section 8.1.1) as well as in the Structured, Social, and Semantic Search project (Section 8.1.2).

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. OAKresearch results directly applicable to this area provide techniques and tools for rich Web content warehouse management. This work will be funded by the ANR ContentCheck project, and a Google Award on Even Thread Extraction. We work in collaboration with Le Monde's "Les Décodeurs" team to investigate these topics.

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. We work on this topic within the 4-year project ODIN started in 2014.

4.4. Hybrid Data Warehousing

Increasingly many modern applications need to exploit data from a variety of formats, including relations, text, trees, graphs etc. The recent development of data management systems aimed at "Big Data", including NoSQL platforms, large-scale distributed systems etc. provides enteprise architects with many systems to chose from. This makes it hard to decide which part of the application data to handle in which system, especially given that each system is best at handling a specific kind of data and a certain class of operations. OAKinvestigates principled techniques for distributing an application's data sources across a variety of systems and data models, based on materialized views. We test our ideas in this area within the Datalyse project.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. Awards

I. Manolescu and X. Tannier (LIMSI) have obtained a Google Computational Research Journalism Award on "Event Thread Extraction for Viewpoint Analysis". The team has also secured an ANR contract on content management techniques applied to computational fact-checking (coordonated by I. Manolescu, to start in 2016) and an ADT engineer has joined the team to work on the same topic.

The best publications of the year appeared in SIGMOD citecamachorodriguez:hal-01178490, PODS [16], PVLDB [29], [8], [26], ICDE [15], [14], and IEEE TKDE [3]. Other highly visible publications appeared in CIDR [9] and CIKM [28], [7].

5.1.2. Inria researcher recruited

M. Thomazo has joined the team as a junior researcher (Inria CR2).

6. New Software and Platforms

6.1. Amada

FUNCTIONAL DESCRIPTION

AMADA is a platform for storing Web data (in particular, XML documents and RDF graphs) based on the Amazon Web Services (AWS) cloud infrastructure. AMADA operates in a Software as a Service (SaaS) approach, allowing users to upload, index, store, and query large volumes of Web data.

- Participants: Jesùs Camacho-Rodriguez, Manolescu Ioana, Dario Colazzo and François Goasdoué
- Contact: Ioana Manolescu
- URL: https://team.inria.fr/oak/projects/amada/

6.2. Clique Square

RDF data management platform based on Hadoop architecture
KEYWORDS: Map-Reduce - Hadoop - RDF - Big data
SCIENTIFIC DESCRIPTION

CliqueSquare is a system for storing and querying large RDF graphs relying on Hadoop's distributed file system (HDFS) and Hadoop's MapReduce open-source implementation. CliqueSquare is equipped with a unique optimization algorithm capable of generating highly parallelizable flat query plans relying on n-ary equality joins. In addition, it provides a novel partitioning and storage scheme that permits first-level joins to be evaluated locally using efficient map-only joins.
FUNCTIONAL DESCRIPTION

RDF (Ressource Description Framework) is the data format for the semantic web. CliqueSquare allows storing and querying very large volumes of RDF data in a massively parralel fashion in a Hadoop cluster. The system uses its own partitioning and storage model for the RDF triples in the cluster.

CliqueSquare evaluates queries expressed in a dialect of the SPARQL query language. It is particularly efficient when processing complex queries, because it is capable of translating them into MapReduce programs guaranteed to have the minimum number of successive jobs. Given the high overhead of a MapReduce job, this advantage is considerable.

- Participants: Ioana Manolescu, Benjamin Djahandideh, Stamatios Zampetakis, Zoi Kaoudi, François Goasdoué and Jorge Arnulfo Quiane Ruiz
- Partners: Université de Rennes 1 - Qatar Computing Research Institute
- Contact: Ioana Manolescu
- URL: https://team.inria.fr/oak/projects/cliquesquare/

6.3. FactMinder

KEYWORDS: Web - Fact-checking - Data Journalism - Open data
FUNCTIONAL DESCRIPTION

FactMinder is a browser extension targeted at online fact checkers and data journalists. It enables users to analyze web pages with entity extractors and create, in a separate panel, views to cross these annotations with background knowledge from trusted XML or RDF sources such as data sets from the Linked Open Data or governmental agencies.

FactMinder is the basis of the ANR project ContentCheck and was awarded a Google Computational Journalism Research Award in June 2015.

- Participants: Ioana Manolescu, Stamatios Zampetakis and François Goasdoué
- Partner: Université Paris-Sud
- Contact: Ioana Manolescu
- URL: https://team.inria.fr/oak/projects/xr-an-xml-rdf-hybrid-model-for-annotated-documents/

6.4. PAXQuery

FUNCTIONAL DESCRIPTION

The PAXQuery engine seamlessly parallelizes the execution of XQuery queries. By applying on-the-fly translation and optimization procedures, PAXQuery runs user queries over massive collections of XML documents in a distributed fashion. PAXQuery runs on top of Apache Flink, a distributed execution platform that relies on the PACT model.

- Participants: Jesùs Camacho-Rodriguez, Ioana Manolescu, Dario Colazzo and Juan Alvaro Munoz Naranjo
- Contact: Ioana Manolescu
- URL: https://team.inria.fr/oak/projects/paxquery/

6.5. RDFSummary

RDF Summary
FUNCTIONAL DESCRIPTION

RDF Summary is a standalone Java software capable of building summaries of RDF graphs. Summaries are compact graphs (typically several orders of magnitude smaller than the original graph), which can be used to get acquainted quickly with a given graph, they can also be used to perform static query analysis, infer certain things about the answer of a query on a graph, just by considering the query and the summary.

- Contact: Sejla Cebiric
- URL: https://team.inria.fr/oak/projects/rdfsummary/

6.6. WaRG

Warehousing RDF Graphs
KEYWORDS: Data mining - Semantic Web - Data management - Decision - Big data
SCIENTIFIC DESCRIPTION

WaRG is a warehouse-style analytics platform on RDF graphs. The tool stores data in kdb+ with a Java frontend based on the Prefuse Visualization toolkit. The novelty of WaRG is to redesign the full stack of Data Warehouse abstractions and tools for heterogeneous, semantics-rich RDF data, this enables a WaRG RDF DW to be an RDF graph itself, heterogeneous and semantics-rich in its turn. Thus, WaRG benefits both from powerful analytics and the rich interoperability and semantic features of Semantic Web databases.
FUNCTIONAL DESCRIPTION

WaRG (Warehousing RDF graph) is an analytical platform specially designed for the analysis of RDF data.

WaRG allows defining RDF analytical schemas, comprising classes and properties interesting for the analysis. The analytical schema can then be materialized, leading to an instance (RDF graph) refined for the needs of the analysis.

The analytical schema can also be automatically built from the input RDF instance. Finally, RDF analytical queries can be specified and lead to RDF analysis cubes.

- Participants: Roatis Alexandra, Ioana Manolescu, Sejla Cebiric and François Goasdoué
- Partners: Université de Rennes 1 - Université Paris-Sud
- Contact: Ioana Manolescu
- URL: https://team.inria.fr/oak/projects/warg/

7. New Results

7.1. Scalable and Expressive Techniques for the Semantic Web

On the topic of efficient query answering methods for semantic-rich RDF data, we have obtained new fundamental results for the RDF Schema ontology language [25] and for a simple DL-Lite dialect [23], [34]; we presented our results in a tutorial at IEEE ICDE [10] and in an invited keynote at SEBD, the Italian Database conference [4]. A demonstration issued from this work was presented at VLDB [26] and at BDA, the French database conference [27].

To help users get acquainted with large and complex RDF graphs, we have started to work on an approach for RDF graph summarization: a graph summary is a smaller RDF graph, often by several orders of magnitude, which preserves the core structural information of the original graph and thus allows to reason about several important graph property on a much more manageable structure. Our first results were presented in [17] and demonstrated at [29] and [30]. These results were also presented in the keynote of the Data Engineering and the Semantic Web workshop [5].

On the related topic of analytical RDF schemas, we have published novel techniques for incrementally computing the result of an RDF analytical query (also known as "RDF cube") out of the result of a previously computed RDF cube [31]. Such computations, commonly known as roll-up, drill-down etc. in the classical relational database setting, require novel solutions for RDF due to the heterogeneity of the graph structure.

7.2. Massively Distributed Data Management Systems

One of the main results of the year is the publication of the full paper [15] and demonstration [14] on CliqueSquare in the highly prestigious IEEE Conference on Data Engineering (ICDE). CliqueSquare has also been released in open source in 2014 (see the Software section). Its main advantage is a novel technique for optimizing conjunctive queries in a massively parallel setting, using n-ary join operators; this allow the optimization algorithm to build plans which are as flat as possible. These results apply beyond the RDF conjunctive query evaluation to the general setting of relational conjunctive query processing in a massively parallel context.

Another crucial result of the year is the publication of the PAXQuery framework for massively processing XML queries based on the Stratosphere (now Apache Flink) platform [3]. We show that our algebra-based approach allows to capture the expressive processing performed by an XQuery query and to compile it efficiently into massively distributed plans which are then evaluated by the Flink platform; this outperforms a set of state-of-the-art approaches for evaluating XQuery queries in a parallel environment. The system was also demonstrated at SIGMOD [11].

7.3. Advanced Algorithms for Data Querying and Transformation

We focused on explaining why some data, so-called missing answers, are not part of the result of a query, even though a developer expects them to be there.

The query-based explanations we return during query analysis serve as the starting point for our query rewriting process. Indeed, knowing the condition combinations pruning data relevant to the missing answers significantly narrows the search space for eligible query rewritings as we can first focus on finding solutions that only affect these query conditions. To further prune the search space, our current solution applies a cost model for rewritings based on several criteria, including edit distance to the original query, or the number of side-effects (tuples additionally appearing in the result of the rewritten query that are not our original missing answers). To select the best solutions w.r.t. the different dimensions of our cost model, we compute and return the skyline over these. We have demonstrated a preliminary version of the proposed algorithn in [8]. This work is reported [7] and in the PhD thesis of K. Tzompanaki [1].

7.4. Social Data Management and Crowdsourcing

Some particular tasks such as annotating data or matching entities have traditionnally been outsourced to human workers for many years. But the last few years have seen the rise of a new research field called crowdsourcing that aims at delegating a wide range of tasks to human workers. Crowd workers tend to make mistakes, so that redundant tasks are typically submitted to mitigate errors. As the crowd is a relatively expansive resource, we have worked on building formal frameworks to improve the efficiency of these processes.

Our research has been focused on two kinds of queries: boolean queries (asking the crowd to identify relevant items in a list, e.g., meals containing a specific ingredient), and ranking queries (asking the crowd to retrieve one or a few preferred items; e.g., ski resorts). We proposed new algorithms and heuristics improving the state of the art for boolean queries, and claimed the first algorithms for ranking queries (more specifically, for top-k and skyline queries) in the comparison framework [16].

We considered top-k query answering in social tagging systems, also known as folksonomies, a problem that requires a significant departure from existing, socially agnostic techniques. In a network-aware context, one can and should exploit the social links, which can indicate how users relate to the seeker and how much weight their tagging actions should have in the result build-up. Beyond explicit social links, we also focus uncovering implicit, potentially richer relationships from user interactions and exploiting them to improve core functionality such as search. Specifically we considered as-you-type search in a social network, where results socially close to the user asking the query are more relevant, and proposed an efficient algorithm presenting, for any (increasingly longer) prefix of the query as the user types it, the k most relevant results [28].

8. Partnerships and Cooperations

8.1. National Initiatives

8.1.1. ANR

Content Management Techniques for Fact-Checking: Models, Algorithms, and Tools (ContentCheck) is a 4-year project starting in January 2016, supported by ANR under DEFI 7 - Société de l'information et de la communication. The project is coordinated by Ioana Manolescu; Bogdan Cautis and Michaël Thomazo also participate. Other partners are U. Rennes 1, INSA Lyon, Le Monde's fact-checking team, and the LIMSI lab of Université Paris Sud. The project aims at establishing fact-checking as a data management problem, and endow it with the appropriate fundamental models, algorithm and tools, validated in interaction with the journalists.

Apprentissage Adaptatif pour le Crowdsourcing Intelligent et l'Accès à l'Information (ALICIA) is a 4-year project, started in February 2014, supported by the ANR CONTINT call. The project is coordinated by Bogdan Cautis, with Nicole Bidoit, and Ioana Manolescu; other partners include LIG (Grenoble) and the Vodkaster company. Its goal is to study models, techniques, and the practical deployment of adaptive learning techniques in user-centric applications, such as social networks and crowdsourcing.

Cloud-Based Organizational Design (CBOD) is a 4-year ANR started in 2014, coordinated by prof. Ahmed Bounfour from UNIV. PARIS-SUD. 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.

Datalyse is funded for 3.5 years as part of the *Investissement d'Avenir - Cloud & Big Data* national program. The project is led by the Grenoble company Eolas, a subsidiary of Business & Decision. It is a collaboration with LIG Grenoble, U. Lille 1, U. Montpellier, and Inria Rhône-Alpes aiming at building scalable and expressive tools for Big Data analytics.

8.1.2. LabEx, IdEx

Structured, Social and Semantic Search is a 3-year project started in October 2013, financed by the *LabEx (Laboratoire d'Excellence)* DIGICOSME. The project aims at developing a data model for rich structured content enriched with semantic annotations and authored in a distributed setting, as well as efficient algorithms for top-k search on such content.

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 in 2014, funded by the Direction Générale de l'Armement, between the SemSoft company, IRISA Rennes and Inria Saclay (OAK). The project aims to develop a complete framework for analytics on Web data, in particular taking into account uncertainty, based on Semantic Web technologies such as RDF.

Google Award I. Manolescu has received a Google Award in collaboration with X. Tannier from LIMSI. The award is given within a call specifically dedicated to computing tools for computational journalism. The project given the award focuses on "Event Thread Extraction for Viewpoint Analysis".

8.2. International Initiatives

8.2.1. Inria International Labs

Inria@SiliconValley

Associate Team involved in the International Lab:

8.2.1.1. OAKSAD

Title: Languages and techniques for efficient large-scale Web data management

International Partner (Institution - Laboratory - Researcher):

University of California, San Diego (United States) - Computer Science and Engineering (CSE) - Alin Deutsch

Start year: 2013

See also: https://team.inria.fr/oak/oaksad/

Data on the Web is increasingly large and complex. The ways to process and share it have also evolved, from the classical scenario where users connect to a database, to today's complex processes whereas data is jointly produced on the Web, disseminated through streams, corroborated and enriched through annotations, and exploited through complex business processes, or workflows. The OAK and San Diego teams work together to devise expressive languages, efficient techniques and scalable platforms for such applications. Our work in 2015 has focused on scalable hybrid stores [9], [18]. The OAKSAD team ended with 2015 but we continue collaborating on this topic.

8.3. International Research Visitors

8.3.1. Visits of International Scientists

- Erietta Liarou, Harvard University, May 2015
- Helena Galhardas, University of Lisbon, March 2015
- Paolo Papotti, Qatar Computing Research Institute, February 2015
- Puya - Hossein Vahabi, Yahoo Labs, January 2015
- Yanlei Diao, University of Massachusetts Amherst, January 2015

8.3.2. Visits to International Teams

8.3.2.1. Sabbatical programme

Bogdan Cautis went on a sabbatical to Hong Kong starting in September 2015, for a duration of one year.

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. General chair, scientific chair

I. Manolescu has been a co-chair of the Digicosme Spring School in Databases (May 2015) http://labex-digicosme.fr/SpringSchool+2015+%28en%29.

9.1.2. Scientific events selection

9.1.2.1. Member of the conference program committees

- I. Manolescu has been a Review Board member of PVLDB 2015 (Experiment and Analysis track), and a PC member for PODS 2015, SIGMOD 2015 (Demonstrations track), BICOD 2015, the Data Engineering and the Semantic Web (DESWeb) workshop 2015 as well as BDA (the French database conference) 2015

9.1.3. Journal

9.1.3.1. Member of the editorial boards

- I. Manolescu has been an associate editor for the ACM Transactions on the Web since september 2015.
- I. Manolescu is a member of the editorial board of the Springer "Data-Centric Systems and Applications" book series

9.1.4. Invited talks

- I. Manolescu gave two keynote talks: on efficient query answering techniques for RDF at the SEBD 2015 conference [4], and on RDF summarization at DESWeb workshop in conjunction with ICDE 2015 [5].

9.1.5. Leadership within the scientific community

N. Bidoit and I. Manolescu are members of the BDA steering committee.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : B. Groz, Introduction to Databases, 28 ETD, L2, Univ. Paris-Sud, France

Licence : B. Groz, Databases, 71 ETD, L3 (and M1), Univ. Paris-Sud, France

Master : B. Groz, Data Warehouses and Olap, 65 ETD, M1 MIAGE, Univ. Paris-Sud, France

Master : I. Manolescu, Architectures for Massively Distributed Data Management, 28 ETD, M2R IAC, Univ. Paris-Sud, France

9.2.2. Supervision

PhD: Katarina Tzompanaki: “Foundations and Algorithms to Compute the Provenance of Missing Data”, defended in December 2015, Melanie Herschel and Nicole Bidoit.

PhD: Stamatis Zampetakis: “Massively Parallel Algorithms for Semantic Web Data”, defended in September 2015, François Goasdoué and Ioana Manolescu.

PhD in progress : Raphael Bonaque: “Structured, Social and Semantic Search”, since October 2013, Bogdan Cautis, François Goasdoué, and Ioana Manolescu.

PhD in progress : Damien Bursztyn: “Scalable Techniques for Web Data Management”, since January 2014, François Goasdoué and Ioana Manolescu.

PhD in progress : Sejla Čebirić: “CloudSelect: Data Mobility Within, Across and Outside Clouds”, since September 2015, A.Bounfour, F. Goasdoué and I. Manolescu.

PhD in progress : Paul Lagrée: “Adaptive Learning for Intelligent Crowdsourcing and Information Access”, since October 2014, Bogdan Cautis.

9.2.3. Juries

Nicole Bidoit has been a member of the PhD commitee of Katarina Tzompanaki.

François Goasdoué has been a member of the PhD committee of Stamatis Zampetakis.

Ioana Manolescu has been a member of the PhD committee of Stamatis Zampetakis, and of the HDR committee of Sarah Cohen-Boulakia (Université de Paris Sud).

9.3. Popularization

- I. Manolescu has given a conférence "Big Data and the Internet" in January in the *UniverCité Ouverte* conference series organized by the city of Gif sur Yvette for the general public .
- R. Bonaque, D. Bursztyn, S. Cebiric and I. Manolescu have presented a game based on RDF graphs and social networks as part of Fête de la Science in Inria Saclay, in October.
- I. Manolescu has presented a vision of data management techniques for journalistic fact-checking at the congress of the french Association of the Information Press Journalists (Association des Journalistes de la Presse d'Information) in October. The same topic has lead to articles in the general press, in Ouest France (http://www.ouest-france.fr/leditiondusoir/data/569/reader/reader.html#!preferred/1/package/569/pub/570/page/8), the canadian Le Devoir (http://www.ledevoir.com/politique/canada/450937/sur-la-piste-du-mensonge), Le Monde (http://data.blog.lemonde.fr/2015/10/23/le-fact-checking-peut-il-sautomatiser/), Rue89 (http://rue89.nouvelobs.com/2015/10/26/algorithmes-antimensonge-fin-bobards-politique-261827), as well as in the Journal du CNRS (https://lejournal.cnrs.fr/articles/un-logiciel-qui-decrypte-la-politique) and Inria's national Web site (http://www.inria.fr/centre/saclay/actualites/un-logiciel-de-fact-checking-pour-comprendre-le-monde-qui-nous-entoure).
- I. Manolescu has presented a database management vision for Data Science in the Big Data Business Convention in November, attended by 600 participants among which half were academics and half from the industry. S. Zampetakis has presented CliqueSquare at the Business Convention.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] K. TZOMPANAKI. *Query Debugging and Fixing to Recover Missing Query Results*, Université Paris Sud - Paris XI, December 2015

[2] S. ZAMPETAKIS. *Scalable algorithms for cloud-based Semantic Web data management*, Université Paris Sud - Paris XI, September 2015, https://tel.archives-ouvertes.fr/tel-01241805

Articles in International Peer-Reviewed Journals

[3] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, I. MANOLESCU. *PAXQuery: Efficient Parallel Processing of Complex XQuery*, in "IEEE Transactions on Knowledge and Data Engineering", July 2015, vol. 27, n^{o} 7, pp. 1977 - 1991 [*DOI :* 10.1109/TKDE.2015.2391110], https://hal.archives-ouvertes.fr/hal-01162929

Invited Conferences

[4] I. MANOLESCU. *Database Optimization Techniques for Semantic Queries*, in "Sistemi Evolutivi di Basi di Dati (SEBD)", Gaeta, Italy, June 2015, https://hal.inria.fr/hal-01179477

[5] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-oriented Summarization of RDF Graphs*, in "Data Engineering Meets Semantic Web Workshop (DESWeb)", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01179484

International Conferences with Proceedings

[6] E. AKBARI-AZIRANI, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Efficient OLAP Operations For RDF Analytics*, in "International Workshop on Data Engineering meets the Semantic Web (DESWeb)", Seoul, South Korea, April 2015 [*DOI :* 10.1109/ICDEW.2015.7129548], https://hal.inria.fr/hal-01187448

[7] N. BIDOIT, M. HERSCHEL, A. TZOMPANAKI. *Efficient Computation of Polynomial Explanations of Why-Not Questions*, in "24th ACM International Conference on Information and Knowledge Management - CIKM 2015", Melbourne, Australia, October 2015 [*DOI :* 10.1145/2806416.2806426], https://hal.archives-ouvertes.fr/hal-01182101

[8] N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *EFQ: Why-Not Answer Polynomials in Action*, in "41st International Conference on Very Large Data Bases - VLDB 2015", Hawai, United States, August 2015, https://hal.archives-ouvertes.fr/hal-01182115

[9] F. BUGIOTTI, D. BURSZTYN, A. DEUTSCH, I. ILEANA, I. MANOLESCU. *Invisible Glue: Scalable Self-Tuning Multi-Stores*, in "Conference on Innovative Data Systems Research (CIDR)", Asilomar, United States, January 2015, https://hal.inria.fr/hal-01087624

[10] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Reasoning on Web Data: Algorithms and Performance*, in "ICDE - 31st International Conference on Data Engineering", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01148500

[11] J. CAMACHO-RODRÍGUEZ, D. COLAZZO, I. MANOLESCU, J. A. M. NARANJO. *PAXQuery: Parallel Analytical XML Processing*, in "ACM SIGMOD International Conference on Management of Data 2015", Melbourne, Victoria, Australia, May 2015, pp. 1117-1122 [*DOI :* 10.1145/2723372.2735374], https://hal.archives-ouvertes.fr/hal-01178490

[12] M. CHEATHAM, Z. DRAGISIC, J. EUZENAT, D. FARIA, A. FERRARA, G. FLOURIS, I. FUNDULAKI, R. GRANADA, V. IVANOVA, E. JIMÉNEZ-RUIZ, P. LAMBRIX, S. MONTANELLI, C. PESQUITA, T. SAVETA, P. SHVAIKO, A. SOLIMANDO, C. TROJAHN DOS SANTOS, O. ZAMAZAL. *Results of the Ontology Alignment Evaluation Initiative 2015*, in "10th ISWC workshop on ontology matching (OM)", Bethlehem, United States, October 2015, pp. 60-115, cheatham2016a, https://hal.archives-ouvertes.fr/hal-01254907

[13] M. COURGEON, T. GIRAUD, M. TARDIEU, A. ROATIS, M. GOUIFFÈS, X. MAÎTRE. *Miroir 3D augmenté par imagerie médicale : la perception de soi en question*, in "27ème conférence francophone sur l'Interaction Homme-Machine", Toulouse, France, October 2015, 2 p. , https://hal.archives-ouvertes.fr/hal-01219988

[14] B. DJAHANDIDEH, F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J.-A. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare in Action: Flat Plans for Massively Parallel RDF Queries*, in "International Conference on Data Engineering", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01108710

[15] F. GOASDOUÉ, Z. KAOUDI, I. MANOLESCU, J.-A. QUIANÉ-RUIZ, S. ZAMPETAKIS. *CliqueSquare: Flat Plans for Massively Parallel RDF Queries*, in "International Conference on Data Engineering", Seoul, South Korea, April 2015, https://hal.inria.fr/hal-01108705

[16] B. GROZ, T. MILO. *Skyline Queries with Noisy Comparisons*, in "ACM SIGMOD/PODS", Melbourne, Australia, May 2015 [*DOI :* 10.1145/2745754.2745775], https://hal.inria.fr/hal-01146568

[17] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-Oriented Summarization of RDF Graphs*, in "Data Science - 30th British International Conference on Databases, BICOD 2015, Edinburgh, UK, July 6-8, 2015, Proceedings", Edinburgh, United Kingdom, July 2015, pp. 87–91, https://hal.inria.fr/hal-01176277

National Conferences with Proceedings

[18] F. BUGIOTTI, D. BURSZTYN, A. DEUTSCH, I. ILEANA, I. MANOLESCU. *Toward Scalable Hybrid Stores*, in "SEBD Italian Symposium on Advanced Database Systems", Gaeta, Italy, June 2015, https://hal.inria.fr/hal-01174301

[19] F. BUGIOTTI, L. CABIBBO, P. ATZENI, R. TORLONE. *How I Learned to Stop Worrying and Love NoSQL Databases*, in "SEBD Italian Symposium on Advanced Database Systems", Gaeta, Italy, June 2015, https://hal.inria.fr/hal-01174303

Conferences without Proceedings

[20] N. BIDOIT, D. COLAZZO, C. SARTIANI, A. SOLIMANDO, F. ULLIANA. *Andromeda: A System for Processing Queries and Updates on Big XML Documents*, in "BigDap 2015 - 2nd International Workshop on Big Data Applications and Principles", Poitiers, France, September 2015, https://hal.archives-ouvertes.fr/hal-01169275

[21] N. BIDOIT, D. COLAZZO, C. SARTIANI, A. SOLIMANDO, F. ULLIANA. *Queries and Updates on Big XML Documents (Extended Abstract)*, in "SEBD 2015 - 23rd Italian Symposium on Advanced Database Systems", GAETA, Italy, June 2015, https://hal.archives-ouvertes.fr/hal-01169269

[22] N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *Efficient computation of polynomial explanations of Why-Not questions*, in "31ème Conférence sur la Gestion de Données — Principes, Technologies et Applications - BDA 2015", Île de Porquerolles, France, September 2015, https://hal.archives-ouvertes.fr/hal-01182104

[23] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Efficient Query Answering in DL-Lite through FOL Reformulation (Extended Abstract)*, in "28th International Workshop on Description Logics", Athens, Greece, June 2015, https://hal.inria.fr/hal-01155715

[24] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Optimizing FOL reducible query answering*, in "BDA'15", Île de Porquerolles, France, September 2015, https://hal.inria.fr/hal-01174300

[25] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Optimizing Reformulation-based Query Answering in RDF*, in "EDBT: 18th International Conference on Extending Database Technology", Brussels, Belgium, March 2015, https://hal.inria.fr/hal-01143068

[26] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Reformulation-based query answering in RDF: alternatives and performance*, in "Very Large Data Bases", Hawaii, United States, August 2015, https://hal.inria.fr/hal-01174298

[27] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Understanding and Improving Reformulation-Based Query Answering Performance in RDF*, in "BDA'15", Île de Porquerolles, France, September 2015, https://hal.inria.fr/hal-01174299

[28] P. LAGRÉE, B. CAUTIS, H. VAHABI. *A Network-Aware Approach for Searching As-You-Type in Social Media*, in "24th ACM International Conference on Information and Knowledge Management - CIKM 2015", Melbourne, Australia, October 2015, https://hal.inria.fr/hal-01181205

[29] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-Oriented Summarization of RDF Graphs*, in "Proceedings of the VLDB Endowment", Kohala Coast, Hawaii, United States, August 2015, vol. 8, n° 12, https://hal.inria.fr/hal-01178140

[30] Š. ČEBIRIĆ, F. GOASDOUÉ, I. MANOLESCU. *Query-Oriented Summarization of RDF Graphs*, in "BDA (Bases de Données Avancées)", Île de Porquerolles, France, September 2015, https://hal.inria.fr/hal-01176301

Research Reports

[31] E. AKBARI-AZIRANI, F. GOASDOUÉ, I. MANOLESCU, A. ROATIS. *Efficient OLAP Operations for RDF Analytics*, OAK team, Inria Saclay ; Inria, January 2015, n° RR-8668, https://hal.inria.fr/hal-01101843

[32] N. BIDOIT, M. HERSCHEL, K. TZOMPANAKI. *Efficiently and Effectively Answering Why-Not Questions based on Provenance Polynomials*, OAK team, Inria Saclay ; Inria, March 2015, n° RR-8697, 25 p. , https://hal.inria.fr/hal-01131561

[33] R. BONAQUE, B. CAUTIS, F. GOASDOUÉ, I. MANOLESCU. *Social, Structured, and Semantic Search*, Inria Saclay Ile de France, October 2015, n° RR-8797, 38 p. , https://hal.inria.fr/hal-01218116

[34] D. BURSZTYN, F. GOASDOUÉ, I. MANOLESCU. *Efficient query answering in the presence of DL-LiteR constraints*, Inria Saclay ; Inria, April 2015, n° RR-8714, https://hal.inria.fr/hal-01143498

Scientific Popularization

[35] S. ROY CHOWDHURY, H. PUROHIT, M. IMRAN. *D-Sieve: A Novel Data Processing Engine for Efficient Handling of Crises-Related Social Messages*, in "Social Web for Disaster Management (SWDM'15)", Florence, Italy, May 2015 [*DOI :* 10.1145/2740908.2741731], https://hal.inria.fr/hal-01136744