# Post-actes BDA 2019 -Gestion de Données Principes Technologies et Applications

Karine Zeitouni, Philippe Lamarre

# Post-actes BDA 2019

—

# Gestion de Données
# Principes Technologies et Applications

Karine Zeitouni[1] and Philippe Lamarre[2]

[1]DAVID - Université de Versailles Saint-Quentin. Karine.Zeitouni@uvsq.fr
[2]LIRIS - INSA Lyon, Université de Lyon. Philippe.Lamarre@insa-lyon.fr

35ème édition
15-18 octobre 2019, Lyon

Actes de la conférence BDA 2019

Conférence soutenue par

Organisation académiques locales

Soutiens

Partenaires industriels

Site de la conférence : `https://bda.liris.cnrs.fr`

# 1 Message des organisateurs

La conférence sur la “Gestion de Données - Principes, Technologies et Applications” BDA est un rendez-vous incontournable de la communauté gestion de données en France. La 35ème édition (BDA 2019) a eu lieu à Lyon, du 15 au 18 octobre 2019, sur le Campus Lyon Tech à la Doua (Laboratoire LIRIS, INSA de Lyon & Université Lyon 1, Villeurbanne). Ces actes regroupent les versions courtes des articles présentés lors de cette conférence, ainsi qu'un résumé des conférences invitées et des tutoriels associés.

La recherche en gestion de données n'a jamais été aussi active, variée, ouverte sur d'autres champs de l'informatique et, au-delà, sur les grands défis des applications modernes. Poursuivant la tradition de rencontres annuelles de la communauté de gestion de données francophone, BDA 2019 a réuni des chercheurs académiques et industriels pour partager leur vision sur les défis et les avancées scientifiques et techniques dans ce domaine en pleine effervescence. BDA 2019 a accueilli 120 participants, parmi lesquels 35 doctorants.

Le programme scientifique a comporté deux conférences invitées, trois tutoriels, une table ronde industrielle, 27 articles de recherche dont 22 longs et 5 courts, 10 démonstrations et 7 articles de doctorants. Depuis plusieurs années, la conférence BDA propose deux catégories d'articles : ceux originaux et les articles déjà acceptés ou publiés dans une conférence internationale de renom. Cela permet de rendre compte globalement des travaux de qualité de la communauté française dans le domaine. De plus, BDA 2019 a récompensé les meilleurs papiers de recherche (deux prix exæquo), une meilleure démonstration, et les meilleures thèses (deux prix thèses exæquo et un accessit).

Nous tenons à remercier tous les auteurs pour leur excellente contribution, la remarquable équipe d'organisation, les membres du comité de programme et le comité de démonstrations, les commissions de prix d'articles de recherche, de démonstration et de thèse. Nos remerciements vont également aux éminents conférenciers invités, Leonid Libkin et Amr El Abbadi, pour l'excellence de leur prestation. Nous avons été honorés de la participation d'industriels à nos côtés, Oracle et DIMO Software, que nous remercions chaleureusement à la fois pour leur participation à la table ronde industrielle et pour leur soutien financier à ces journées. Nous remercions aussi nos autres soutiens et sponsors sans qui cette manifestation n'aurait pas pu avoir lieu. Au niveau local nous remercions les organisation académiques : LIRIS, INSA Lyon, Université de Lyon 1 ; et au niveau régional et national : Grand Lyon, INRIA, CNRS, SIF.

Karine Zeitouni et Philippe Lamarre

# 2 Comités BDA

**Président des Journées**

**Mohand-Saïd Hacid**
LIRIS - Université Lyon 1

**Président du Comité de Pilotage BDA**

**Patrick Valduriez**
Inria & LIRMM, Univ. Montpellier

**Comité de programme**

**Karine Zeitouni**
DAVID - Université de Versailles Saint-Quentin

- Reza Akbarinia (Inria & LIRMM, Univ. Montpellier)
- Sihem Amer-Yahia (LIG - CNRS, Grenoble)
- Sabeur Aridhi (CNRS - Inria Lorraine)
- Ladjel Bellatreche (LIAS - ENSMA, Poitiers)
- Patrice Bellot (LSIS - Univ. Aix-Marseille)
- Nicole Bidoit-Tollu (LRI - Univ. Paris-Sud)
- Omar Boucelma (LSIS - Univ. Aix-Marseille)
- Emmanuel Bruno (LSIS - Univ. Toulon)
- Dario Colazzo (LAMSADE - Univ. Paris-Dauphine)
- Camélia Constantin (LIP6 - Sorbonne Univ.)
- Emmanuel Coquery (LIRIS - Univ. Lyon 1)
- Alexandru Costan (IRISA - Inria Rennes Bretagne-Atlantique)
- Laurent d'Orazio (IRISA - Univ. Rennes 1)
- Cedric Du-Mouza (CEDRIC - CNAM)
- Pierre Geneves (LIG - Inria Grenoble Rhône-Alpes)
- David Gross-Amblard (IRISA, Univ. Rennes 1)
- Zoubida Kedad (DAVID - Univ. Versailles Saint-Quentin)
- Anne Laurent (LIRMM - Univ. Montpellier)
- Sofian Maabout (LABRI - Univ. Bordeaux)
- Ioana Manolescu (Inria Saclay)
- Benjamin Nguyen (LIFO - INSA Centre Val de Loire)
- Esther Pacitti (Inria & LIRMM, Univ. Montpellier)
- Themis Palpanas (LIPAD - Univ. Paris Descartes)
- Paolo Papotti (Eurecom, Sophia Antipolis)
- Jean-Marc Petit (LIRIS - INSA-Lyon)
- Jorge-Arnulfo Quiané-Ruiz (QCRI, Qatar)

- Fatiha Sais (LRI - Univ. Paris Sud)
- Guillaume Scerri (Inria & DAVID - Univ. Versailles Saint-Quentin)
- Pierre Sens (Inria & LIP6 - Sorbonne Univ.)
- Hala Skaf-Molli (LS2N - Univ. Nantes)
- Alexandre Termier (IRISA - Univ. Rennes 1)
- Olivier Teste (IRIT - Univ. Toulouse)
- Farouk Toumani (LIMOS - Univ. Auvergne)
- Genoveva Vargas-Solar (LIG & LAFMIA - CNRS)
- Dan Vodislav (ETIS - Univ. Cergy-Pontoise)
- Agnes Voisard (FU & Fraunhofer FOKUS)
- Shaoyi Yin (IRIT - Univ. Toulouse)

### Comité de démonstration

**Yehia Taher**
DAVID - Université de Versailles Saint-Quentin

- Antoine Amarilli (DIG - Télécom ParisTech)
- Mahmoud Barhamgi (LIRIS - Univ. Lyon 1)
- Mohammad Dbouk (L'ARiCoD - Univ. Libanaise)
- Walid Gaaloul (SAMOVAR - Télécom SudParis)
- Luis Galarraga del Prado (Inria Rennes Bretagne-Atlantique)
- Soror Sahri (LIPADE - Univ. Paris Descartes)
- Yacine Sam (LIFAT - Univ. Tours)
- Iulian Sandu Popa (Inria & DAVID - Univ. Versailles Saint-Quentin)
- Danai Symeonidou (INRA Montpellier)
- Nicolas Travers (DVRC - ESILV, Pôle Universitaire Léonard de Vinci)
- Katerina Tzompanaki (ETIS - Univ. Cergy-Pontoise)

### Comité d'organisation

**Philippe Lamarre**
LIRIS - INSA-Lyon

- Sylvie Cazalens (LIRIS - INSA-Lyon)
- Emmanuel Coquery (LIRIS - Univ. Lyon 1)
- Pierre Genevès (LIG - Inria Grenoble Rhône-Alpes)
- Yann Gripay (LIRIS - INSA-Lyon)
- Jean-Marc Petit (LIRIS - INSA-Lyon)
- Marian Scuturici (LIRIS - INSA-Lyon)
- Romuald Thion (LIRIS - Univ. Lyon 1)
- Guillaume Beslon (LIRIS - INSA-Lyon)

# Table des matières

# 3 Conférences invitées

## 3.1 A Case against SQL (if you need one)

**Leonid Libkin**
**University of Edinburgh (Scotland, United Kingdom)**

This talk grew out of our work on handling nulls in SQL - and in the process our attempt to understand the semantics of SQL. It led to quite a few surprises, reviewed here mainly as a collection of anecdotes, of several kinds :

Extremely simple looking SQL queries that, the more experience you have with SQL, the less clue you have what about their behavior (and for a good reason). Comparison of ostensibly Standard-compliant DBMSs on queries written according to the Standard (moral of the story : don't change your RDBMS). Errors you get when you handle nulls (executive summary : you tell me how many errors in a single query you want, and I give you that query). I shall also talk about what we do to address these issues, for example, providing a formal semantics of actual SQL, without any shortcuts, and understanding the necessity of SQL design choices, such as the much criticized 3-valued logic.

**Prof. Leonid Libkin** is Professor of Foundations of Data Management in the School of Informatics at the University of Edinburgh. He was previously a Professor at the University of Toronto and a member of research staff at Bell Laboratories in Murray Hill. He received his PhD from the University of Pennsylvania in 1994. His main research interests are in the areas of data management and applications of logic in computer science. He has written five books and over 200 technical papers. His awards include a Marie Curie Chair Award, a Royal Society Wolfson Research Merit Award, and six Best Paper Awards. He has chaired programme committees of major database conferences (ACM PODS, ICDT) and was the conference chair of the 2010 Federated Logic Conference. He has given many invited conference talks and has served on multiple program committees and editorial boards. He is an ACM fellow, a fellow of the Royal Society of Edinburgh, and a member of Academia Europaea

## 3.2 A Wakeup Call : Databases in an Untrusted Universe

**Amr El Abbadi**
**University of California (Santa Barbara, USA)**

Once upon a time databases were structured, one size fit all and they resided on machines that were trustworthy and even when they failed, they simply crashed. This era has come and gone as eloquently stated by Mike Stonebraker. We now have key-value stores, graph databases, text databases, and a myriad of unstructured data repositories. However, we, as a database community still cling to our 20th century belief that databases always reside on trustworthy, honest servers. This notion has been challenged and abandoned by many other Computer Science communities, most notably the security and the distributed systems communities.

The rise of the cloud computing paradigm as well as the rapid popularity of blockchains demand a rethinking of our naïve, comfortable beliefs in an ideal benign infrastructure. In the cloud, clients store their sensitive data in remote servers owned and operated by cloud providers. The Security and Crypto Communities have made significant inroads to protect both data and access privacy from malicious untrusted storage providers using encryption and oblivious data stores. The Distributed Systems and the Systems Communities have developed consensus protocols to ensure the fault-tolerant maintenance of data residing on untrusted, malicious infrastructure. However, these solutions face significant scalability and performance challenges when incorporated in large scale data repositories. Novel da-

tabase design needs to directly address the natural tension between performance, fault-tolerance and trustworthiness. This is a perfect setting for the database community to lead and guide.

In this talk, I will discuss the state of the art in terms of data management in malicious, untrusted settings, its limitations and potential approaches to mitigate these shortcomings. As examples, I will use cloud and distributed databases that reside on untrustworthy malicious infrastructure and discuss specific approachs for standard database problems like commitment and replication. I will also explore blockchains, which can be viewed as asset management databases in untrusted infrastructures. In this context, I will discuss recent approaches to improve transaction throughput in blockchains, as well as recent solutions to achieve atomic commitment among multiple not trusting blockchains.

**Prof. Amr El Abbadi** is a Professor of Computer Science at the University of California, Santa Barbara. He received his B. Eng. from Alexandria University, Egypt, and his Ph.D. from Cornell University. His research interests are in the fields of fault-tolerant distributed systems and databases, focusing recently on Cloud data management and blockchain based systems. Prof. El Abbadi is an ACM Fellow, AAAS Fellow, and IEEE Fellow. He was Chair of the Computer Science Department at UCSB from 2007 to 2011. He has served as a journal editor for several database journals, including, The VLDB Journal, IEEE Transactions on Computers and The Computer Journal. He has been Program Chair for multiple database and distributed systems conferences. He currently serves on the executive committee of the IEEE Technical Committee on Data Engineering (TCDE) and was a board member of the VLDB Endowment from 2002 to 2008. In 2007, Prof. El Abbadi received the UCSB Senate Outstanding Mentorship Award for his excellence in mentoring graduate students. In 2013, his student, Sudipto Das received the SIGMOD Jim Gray Doctoral Dissertation Award. Prof. El Abbadi is also a co-recipient of the Test of Time Award at EDBT/ICDT 2015. He has published over 300 articles in databases and distributed systems and has supervised over 35 PhD students

# 4 Tutoriels

## 4.1 Data Pipelines for User Group Analytics

**Behrooz Omidvar-Tehrani** (University of Grenoble Alpes — Grenoble, France) and **Sihem Amer-Yahia** (University of Grenoble Alpes & CNRS — Grenoble, France)

User data is becoming increasingly available in various domains ranging from the social Web to electronic patient health records (EHRs). User data is characterized by a combination of demographics (e.g., age, gender, life status) and user actions (e.g., posting a tweet, following a diet). Domain experts rely on user data to conduct large-scale population studies. Information consumers, on the other hand, rely on user data for routine tasks such as finding a book club and getting advice from look-alike patients. User data analytics is usually based on identifying group-level behaviors such as "teenage females who watch Titanic" and "old male patients in Paris who suffer from Bronchitis." In this tutorial, we review data pipelines for User Group Analytics (UGA). These pipelines admit raw user data as input and return insights in the form of user groups. We review research on UGA pipelines and discuss approaches and open challenges for discovering, exploring, and visualizing user groups. Throughout the tutorial, we will illustrate examples in two key domains : "the social Web" and "health-care".

## 4.2 Schemas and Types for JSON Data : From Theory to Practice

**Mohamed-Amine Baazizi** (LIP6, Sorbonne Université — Paris, France), **Dario Colazzo** (Université Paris-Dauphine — Paris, France), **Giorgio Ghelli** (Università di Pisa — Pisa, Italia) and **Carlo Sartiani** (Università della Basilicata (Potenza, Italia).

The last few years have seen the fast and ubiquitous diffusion of JSON as one of the most widely used formats for publishing and interchanging data, as it combines the flexibility of semistructured data models with well-known data structures like records and arrays. The user willing to effectively manage JSON data collections can rely on several schema languages, like JSON Schema, JSound, and Joi, as well as on the type abstractions offered by modern programming and scripting languages like Swift or TypeScript. The main aim of this tutorial is to provide the audience (both researchers and practitioners) with the basic notions for enjoying all the benefits that schema and types can offer while processing and manipulating JSON data. This tutorial focuses on four main aspects of the relation between JSON and schemas : (1) we survey existing schema language proposals and discuss their prominent features ; (2) we analyze tools that can infer schemas from data, or that exploit schema information for improving data parsing and management ; and (3) we discuss some open research challenges and opportunities related to JSON data.

## 4.3 Database and Distributed Computing Foundations of Blockchains

**Amr El Abbadi** (University of California — Santa Barbara, USA)

The uprise of Bitcoin and other peer-to-peer cryptocurrencies has opened many interesting and challenging problems in cryptography, distributed systems, and databases. The main underlying data structure is blockchain, a scalable fully replicated structure that is shared among all participants and guarantees a consistent view of all user transactions by all participants in the system. In this tutorial, we discuss the basic protocols used in blockchain, and elaborate on its main advantages and limitations. To overcome these limitations, we provide the necessary distributed systems background in managing large scale fully replicated ledgers, using Byzantine Agreement protocols to solve the consensus problem. Finally, we expound on some of the most recent proposals to design scalable and efficient blockchains in both permissionless and permissioned settings. The focus of the tutorial is on the distributed systems and database aspects of the recent innovations in blockchains

# 5 Résumés des articles

# Medical Cohort Analysis: Representation and Exploration

Behrooz Omidvar-Tehrani
NAVER LABS Europe
firstname.lastname@naverlabs.com

Sihem Amer-Yahia
CNRS, University of Grenoble Alpes
firstname.lastname@univ-grenoble-alpes.fr

Laks V.S. Lakshmanan
University of British Columbia
firstname@cs.ubc.ca

With the increase in health-care data in various sectors (e.g., prognoses, treatments, hospitalizations and compliances), medical experts require effective analysis methods to understand the evolution of their patients' health. Medical cohort analysis exhibits the collective behavior of patients, providing insights on the evolution of their health conditions and their reaction to treatments. Cohort analysis serves various goals such as augmenting treatment effectiveness, patient satisfaction, and health-care revenue [1].

In medical cohort analysis, experts seek answers to three recurring questions: *"what will happen next?"*, *"what has happened?"*, and *"what happened to similar cohorts?"* The first question relates to *predicting* the future status of patients in a cohort based on their past. The second question relates to *finding and conveying a representation* of patient data in a human-understandable way. The third question relates to *exploring* the data to suggest similar cohorts to experts and help them find differences and contextualize decisions. Existing work mostly focused on classification and prediction, e.g., using association rule mining to classify cohorts [2], or deep neural network to predict disease progression and future risks [3]. Representation and exploration of health-care data in general, and of cohorts in particular, have received little attention. As an example, representing cohorts of patients with respiratory problems helps experts to find an answer for questions like *"which sequence of treatments is the most relevant to an outcome, e.g., death?"*, *"what changes in Body Mass Index (abbr., BMI) lead to death?"*, *"which treatment is employed right after admission which kept cohort members alive for a longer period?"* Also, exploring cohorts of those patients helps experts to compare, e.g., patients in urban and rural regions and verify the effect of air pollution on the way the collective behavior of those sub-cohorts changes. Hence, in this paper, we focus on (*i*) providing a readable and succinct representation of medical cohorts, and (*ii*) enabling cohort exploration as the action of finding similar cohorts.

**Representation** of a cohort helps understand which sequence of diseases and treatments lead to an outcome, e.g., death. The aggregation of demographics (e.g., age, gender, occupation) and medical actions (e.g., diseases, prognoses, treatments, marker observations) forms cohorts such as *"middle-age females in Paris with Bronchitis"* and *"old male patients with a sudden BMI decrease during hospitalization."* Obtaining a readable and succinct representation of a cohort is challenging because cohorts often consist of hundreds of patients whose medical actions are of various types and occur at different points in time. Statistical methods provide numerical representations (e.g., mortality rate, average duration of treatments, BMI variations), but their scope is often limited to one aspect of the cohort's status [4]. Similarly, while graph summarization [5] and process mining [6] can be used to aggregate health trajectories of patients in a cohort, the final representation suffers from "imprecision", because the flow of transitions in the cohort is not preserved. Visual Analytics (such as in EventFlow [7]) is helpful, but visual views get cluttered easily when dealing with heterogeneous cohorts, the so-called Confetti effect. In this paper, we extend the Needleman-Wunsch algorithm for sequence matching to handle temporal sequences [8]. Our algorithm aggregates the health trajectories of patients in a cohort into a succinct representation that captures the flow of medical events. Since sequence matching has a quadratic time complexity, we introduce "trajectory families" based on $k$-medoids to improve response time.

**Exploration** refers to navigating in health data to gain an understanding of how different cohorts evolve over time. We define exploration as finding cohorts similar to a given cohort. This is challenging because the potential number of similar cohorts is huge. With only 20 demographic attribute values, $2^{20}$ potential cohorts can be built. Note that real health-care data often contains over 5000 attribute values. We propose a two-staged approach based on generating candidate cohorts and then computing their similarity to the given cohort. In the first stage, we draw inspiration from observational medical studies [9] and limit the set of candidates to "contrast cohorts", those that differ by only one attribute value from the initial cohort. For instance, for exploring female patients in Paris, one candidate cohort is females outside Paris. To build contrast cohorts, we compute a "contribution ratio" for each attribute that reflects the impact of changing a value of that attribute on the cohort's membership. It is intuitive to build contrast cohorts using attributes with a small contribution ratio to ensure they are both similar and different-enough from the input cohort. Once identified, the similarity of each contrast cohort to the given cohort is computed and used to return candidate cohorts to explore. To speed up cohort similarity computation, we use "event sets" in the same spirit as the double dictionary encoding proposed for keyword search [10].

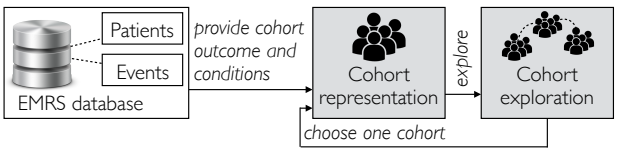


**Figure 1:** Core **framework.**

We build Core, a data-driven framework for **CO**hort **R**epresentation and **E**xploration (see Figure 1 for the overall architecture). To test the efficiency and usefulness of Core, we run experiments on

a real health-care data with a focus on respiratory problems. We study the benefits of trajectory families and event sets in providing interactive performance. In a user study with medical experts, we show that Core reduces time-to-insight from hours to seconds and helps medical experts find better insights when compared to alternative approaches such as SQL querying and Visual Analytics.

## ACKNOWLEDGMENT

The authors would like to thank Jean-Louis Pépin, Pierre Hainaut, Jean-Christian Borel, Sebastien Bailly, Nathanaël Lemonnier and Son T. Mai for their valuable remarks during the course of this work. This work is supported by CDP LIFE project under grant C7H-ID16-PR4-LIFELIG.

## REFERENCES

[1] A Munshi, V Sharma, and S Sharma. Lessons learned from cohort studies, and hospital-based studies and their implications in precision medicine. In *Progress and Challenges in Precision Medicine*. Elsevier, 2017.

[2] Susan Rea Welch and Stanley M Huff. Cohort amplification: An associative classification framework for identification of disease cohorts in the electronic health record. In *Annual Symposium Proceedings*. American Medical Informatics Association, 2010.

[3] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69:218–229, 2017.

[4] Aaron Heuser, Minh Huynh, and Joshua C Chang. Empirical process-based large sample properties of the area bounded by cohort-weighted kaplan meier curves. *arXiv preprint arXiv:1701.02424*, 2017.

[5] Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. Graph summarization methods and applications: A survey. *ACM Computing Surveys*, 2018.

[6] Arik Senderovich, Matthias Weidlich, and Avigdor Gal. Temporal network representation of event logs for improved performance modelling in business processes. In *BPM*, 2017.

[7] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. Temporal event sequence simplification. *Transactions on visualization and computer graphics*, 2013.

[8] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[9] Erik Von Elm, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, Jan P Vandenbroucke, Strobe Initiative, et al. The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *PLoS medicine*, 4(10):e296, 2007.

[10] Alexander Hall, Olaf Bachmann, Robert Büssow, Silviu Gănceanu, and Marc Nunkesser. Processing a trillion cells per mouse click. *Proceedings of the VLDB Endowment*, 5(11):1436–1446, 2012.

# Reformulation-based query answering for RDF graphs with RDFS ontologies

Maxime Buron
Inria
maxime.buron@inria.fr

François Goasdoué
U. Rennes & Inria
fg@irisa.fr

Ioana Manolescu
Inria
ioana.manolescu@inria.fr

Marie-Laure Mugnier
U. Montpellier & Inria
marie-laure.mugnier@lirmm.fr

## ABSTRACT

Query answering in RDF knowledge bases has traditionally been performed either through graph saturation, i.e., 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 [13], 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 this paper, we address the general query answering problem by reducing it, through a pre-query reformulation step, to that solved by the query reformulation technique of [13]. We also report on experiments demonstrating the low cost of our reformulation algorithm.

**ACM Reference Format:**
Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. 2019. Reformulation-based query answering for RDF graphs with RDFS ontologies. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

# SaGe : Préemption Web pour les services publics d'évaluation de requêtes SPARQL*

Thomas Minier
LS2N, Université de Nantes
Nantes, France
thomas.minier@univ-nantes.fr

Hala Skaf-Molli
LS2N, Université de Nantes
Nantes, France
hala.skaf@univ-nantes.fr

Pascal Molli
LS2N, Université de Nantes
Nantes, France
pascal.molli@univ-nantes.fr

## CCS CONCEPTS

• **Information systems → Data management systems**; *Database query processing*; **Resource Description Framework (RDF)**;

## KEYWORDS

Web des données · Web Sémantique · Évaluation de requêtes SPARQL

## 1 INTRODUCTION

**Motivations :** Suivant les principes du Linked Open Data (LOD), les fournisseurs de données hébergent publiquement des millions de triples au format RDF [3, 7]. Cependant, fournir un service public qui permet à n'importe qui d'exécuter n'importe quelle requête SPARQL sur ces données est toujours un problème ouvert. Comme ces services sont soumis à une charge imprévisible de requêtes, le défi est d'assurer qu'ils demeurent *stables* malgré des variations en termes de taux d'arrivées des requêtes et des ressources nécessaires à leur évaluation.

Pour résoudre ce problème, la plupart des fournisseurs de données appliquent une politique d'utilisation équitable des serveurs basée sur des *quotas* [1]. Selon les administrateurs de DBpedia, « A Fair Use Policy is in place in order to provide a stable and responsive endpoint for the community. » [1] En conséquence, le SPARQL endpoint publique de DBpedia interrompt l'exécution des requêtes SPARQL durant plus de 60 secondes ou renvoyant plus de 10000 résultats, avec une limite de 50 connexions concurrentes et 100 requêtes HTTP par seconde par adresse IP.

Ces quotas ont pour objectif de permettre un partage équitable des ressources du serveur Web entre les différents clients. Les quotas sur les taux de communications limitent les taux d'arrivées de requêtes, tandis ceux sur le nombre de résultats empêchent une unique requête de monopoliser toutes les ressources du serveur. Les quotas sur la durée d'exécution visent à empêcher les *effets convois* [4], *c.a.d*, une requête longue à exécuter bloque l'évaluation des autres. Le principal défaut de cette politique est qu'elle empêche les requêtes interrompues de délivrer des résultats complets. Cela constitue une limite sérieuse pour les utilisateurs du LOD, qui peuvent vouloir exécuter des requêtes longues [6].

**État de l'art :** Les approches existantes résolvent ce problème en décomposant les requêtes SPARQL en un ensemble de sous-requêtes qui peuvent être évaluées avec des résultats complets tout en respectant les quotas du serveur [2]. En revanche, trouver cette décomposition est complexe dans le cas général, car les quotas sont différents d'un serveur à l'autre [2]. L'approche Linked Data Fragments (LDF) [5, 8] résout ce problème en restreignant l'interface du serveur à des opérations SPARQL qui peuvent être évaluées sans générer d'effets convois. Par exemple, dans l'approche Triple Pattern Fragments (TPF) [8], un serveur TPF permet uniquement l'évaluation paginée de triple patterns. En contrepartie, les approches basées sur LDF génèrent un large nombre de sous-requêtes et transfèrent une quantité importante de résultats intermédiaires.

**Contributions :** Nous pensons que le problème lié aux quotas ne réside pas dans l'interruption des requêtes, mais dans l'impossibilité pour les clients de *reprendre leur exécution* après interruption. Néanmoins, il n'existe pas de modèle de préemption pour le Web qui permet la suspension et la reprise de l'exécution de requêtes SPARQL. Dans cet article, nous proposons SaGe, un moteur d'évaluation de requêtes SPARQL basé sur la *préemption Web*. Elle permet à un serveur Web de suspendre une requête en cours d'exécution après un certain temps, puis de reprendre son exécution ultérieurement. Une fois suspendue, l'état sauvegardé d'une requête est retourné au client, qui peut reprendre son exécution en renvoyant l'état au serveur. La préemption Web engendre des coûts supplémentaires pour le serveur Web, qui doit suspendre la requête courante puis reprendre l'exécution de la suivante. En conséquences, le problème scientifique majeur est de maintenir ce surcoût marginal, quelle que soit la requête, afin d'assurer une exécution performante. Nos contributions sont les suivantes :

- Nous formalisons le modèle de *préemption Web* qui permet de suspendre et de reprendre l'exécution des requêtes SPARQL. Nous définissons ainsi un ensemble d'opérateurs d'exécution préemptifs, dont les complexités d'arrêt et de reprise sont bornées en temps et en espace.
- Nous proposons SaGe, un moteur d'évaluation de requêtes SPARQL composé d'un serveur Web préemptif et d'un client Web intelligent, qui permet l'évaluation de n'importe quelle requête SPARQL en suivant le modèle de la Préemption Web.
- Nous comparons les performances du moteur de requêtes SaGe avec les approches existantes. Nos résultats expérimentaux démontrent que SaGe surpasse de plusieurs ordres de grandeurs les approches existantes en termes de temps moyen d'exécution des requêtes et de temps d'obtention des premiers résultats.

*. Cet article a été publié dans les actes de the 2019 World Wide Web Conference (WWW'19) https://doi.org/10.1145/3308558.3313652

1. http://wiki.dbpedia.org/public-sparql-endpoint

## ACKNOWLEDGMENTS

Le travail de T. Minier a été partiellement financé par le projet FaBuLA, qui fait partie du programme AtlanSTIC 2020.

## RÉFÉRENCES

[1] Carlos Buil Aranda, Aidan Hogan, Jürgen Umbrich, and Pierre-Yves Vandenbussche. 2013. SPARQL Web-Querying Infrastructure : Ready for Action ?. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II (Lecture Notes in Computer Science)*, Vol. 8219. Springer, 277–293. https://doi.org/10.1007/978-3-642-41338-4_18

[2] Carlos Buil Aranda, Axel Polleres, and Jürgen Umbrich. 2014. Strategies for Executing Federated Queries in SPARQL1.1. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II (Lecture Notes in Computer Science)*, Vol. 8797. Springer, 390–405. https://doi.org/10.1007/978-3-319-11915-1_25

[3] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.* 5, 3 (2009), 1–22. https://doi.org/10.4018/jswis.2009081901

[4] Mike W. Blasgen, Jim Gray, Michael F. Mitoma, and Thomas G. Price. 1979. The Convoy Phenomenon. *Operating Systems Review* 13, 2 (1979), 20–25. https://doi.org/10.1145/850657.850659

[5] Olaf Hartig, Ian Letter, and Jorge Pérez. 2017. A Formal Framework for Comparing Linked Data Fragments. In *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I (Lecture Notes in Computer Science)*, Vol. 10587. Springer, 364–382. https://doi.org/10.1007/978-3-319-68288-4_22

[6] Axel Polleres, Maulik R. Kamdar, Javier D. Fernández, Tania Tudorache, and Mark A. Musen. 2018. A More Decentralized Vision for Linked Data. In *Proceedings of the 2nd Workshop on Decentralizing the Semantic Web co-located with the 17th International Semantic Web Conference, DeSemWeb@ISWC 2018, Monterey, California, USA, October 8, 2018.*

[7] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. 2014. Adoption of the Linked Data Best Practices in Different Topical Domains. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I (Lecture Notes in Computer Science)*, Vol. 8796. Springer, 245–260. https://doi.org/10.1007/978-3-319-11964-9_16

[8] Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert. 2016. Triple Pattern Fragments : A low-cost knowledge graph interface for the Web. *J. Web Sem.* 37-38 (2016), 184–206. https://doi.org/10.1016/j.websem.2016.03.003

# Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue

Rana Alotaibi
UC San Diego
and KACST
ralotaib@eng.ucsd.edu

Damian Bursztyn*
Thales Digital Factory
dbursztyn@gmail.com

Alin Deutsch
UC San Diego
deutsch@cs.ucsd.edu

Ioana Manolescu
Inria
LIX (UMR 7161, CNRS and
Ecole polytechnique)
ioana.manolescu@inria.fr

Stamatis Zampetakis*
TIBCO Orchestra Networks
zabetak@gmail.com

**ABSTRACT**

Big data applications routinely involve diverse datasets: relations flat or nested, complex-structure graphs, documents, poorly struc-tured logs, or even text data. To handle the data, application de-signers usually rely on several data stores used side-by-side, each capable of handling one or a few data models, and each very effi-cient 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.

We present ESTOCADA, a novel approach connecting applica-tions to the potentially heterogeneous systems where their input data resides. ESTOCADA can be used in a polystore setting to trans-parently 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 rewrit-ing under constraints, which we use to describe the various data models and stored data. Our experiments illustrate the significant performance gains achieved by ESTOCADA.

*Work done while the author was a PhD student working at I nria, France.

# Influence Maximization in a Real-Time Bidding Environment

David Dupuis
Kwanko & Léonard de Vinci Pôle Universitaire, Research Center, CNAM
Paris, France
david.dupuis@devinci.fr

Cédric du Mouza
CNAM
Paris, France
cedric.dumouza@cnam.fr

Nicolas Travers
Léonard de Vinci Pôle Universitaire, Research Center
Paris, France
nicolas.travers@devinci.fr

Gaël Chareyron
Léonard de Vinci Pôle Universitaire, Research Center
Paris, France
gael.chareyron@devinci.fr

## ABSTRACT

Influence Maximization (IM) is a well studied maximum coverage problem, which consists in finding in a network the top-k influencers who will maximize the diffusion of a piece of information. It is commonly associated with basic online advertising strategies. However, today, the exponential growth of online advertising is due to Real-Time Bidding (RTB) which allows advertising agen-cies to target specific users with specific ads on specific webpages. It requires complex ad placement decisions in real-time to face a high-speed stream of online users. In order to stay relevant, the IM problem should be updated to answer RTB needs. While most tra-ditional IM methods generate a static set of top-k influencers, they do not deal with the topic of influence maximization in a real-time bidding environment which requires dynamic influence targeting.

This paper studies this topic and proposes RTIM, the first IM al-gorithm capable of taking influence maximization decisions within a real-time bidding environment. We also analyze influence scores of users in several social networks in order to show their impact in our approach. Finally, we offer a thorough experimental process to compare static versus dynamic IM solutions *wrt.* influence scores.

## KEYWORDS

Real-Time Bidding, Influence Maximization, Social Network

**This article has been published in the WISE'19 conference**

# SEP2P: Secure and Efficient P2P Personal Data Processing

Julien Loudet[1,2,3]
[1]Cozy Cloud, France
contact@cozycloud.cc

Iulian Sandu-Popa[3,2]
[2]INRIA Saclay, France
<fname.lname>@inria.fr

Luc Bouganim[2,3]
[3]University of Versailles, France
<fname.lname>@uvsq.fr

Le temps de la réappropriation de nos données numériques personnelles est arrivé. En effet, grâce aux initiatives soutenues par différents gouvernements — *BlueButton* [4] et *GreenButton* aux États-Unis, *MesInfos* [7] en France ou *Midata* [11] au Royaume-Uni — ainsi qu'à de nouvelles réglementations — comme le Réglement Général sur la Protection des Données [13] récemment voté en Europe — les utilisateurs peuvent demander aux entreprises et agences gouvernementales une copie de leurs données. En parallèle de cela, de plus en plus de *Systèmes de Gestion de Données Personnelles* (SGDP) voient le jour, que ce soit dans l'industrie [5, 12, 15] ou dans le milieu académique [2]. Leur objectif est d'offrir une plateforme permettant de facilement conserver et gérer, à un même endroit, les données produites par les appareils des utilisateurs (par exemple, issues de la mesure de soi ou de la maison connectée) et celles qui résultent de leurs interactions avec des services (par exemple sur les réseaux sociaux ou auprès de leur banque). Les utilisateurs peuvent ensuite profiter des capacités offertes par leur SGDP pour effectuer divers calculs sur leurs données dans leur intérêt ou dans l'intérêt d'une communauté. Le paradigme, nouveau, du SGDP promet dès lors de créer de nouveaux usages.

Nous pouvons notamment considérer trois cas d'usage emblématiques d'applications distribuées qui se basent sur de larges communautés d'utilisateurs et qui pourraient grandement bénéficier de ce paradigme : (1) des applications de mesure participatives [16], où les utilisateurs produisent des mesures géo-localisées (par exemple, le trafic automobile, la qualité de l'air) et calculent des agrégations spatiales dans l'intérêt de toute la communauté ; (2) des applications de diffusion [18] qui se basent soit sur une souscription soit sur un profil pour transmettre une information et où le SGDP fournit les préférences ou le profil de l'utilisateur pour que celui-ci ne reçoive que des informations pertinentes ; et (3) des requêtes distribuées sur les données personnelles d'un large ensemble d'individus [17], où les utilisateurs contribuent avec leurs données personnelles et formulent des requêtes sur l'ensemble des contributions (par exemple, pour calculer des recommandations ou faire des études participatives).

Cependant, ces perspectives intéressantes ne devraient pas éclipser les problèmes de sécurité soulevés par le paradigme du SGDP. En effet, chaque SGDP contient potentiellement toute la vie numérique de son détenteur, ce qui augmente proportionnellement l'impact d'une fuite. Ainsi, centraliser l'ensemble des données des utilisateurs sur de puissants serveurs est risqué puisque ces derniers deviennent dès lors les cibles privilégiées des attaquants : d'énorme quantité de données appartenant à des millions d'individus pourraient fuiter, comme illustré par les récentes attaques informatiques. De surcroît, une solution centralisée serait antinomique dans ce contexte puisque les données sont naturellement distribuées chez les utilisateurs [9].

De récents travaux [2, 6, 10, 15] proposent de laisser les données des utilisateurs sur des plateformes de confiance qui sont directement sous le contrôle de ces derniers. De telle plateformes peuvent être construites grâce à la combinaison de : (1) un environnement d'exécution de confiance (par exemple, du matériel sécurisé comme des « cartes à puce » [1] ou des microcontrôleurs sécurisés [2, 3, 10] comme ARM TrustZone [8] et Intel SGX [14]) ; et (2) des logiciels dédiés. Dans cet article, nous proposons de suivre cette approche et considérons qu'un SGDP consiste en un matériel personnel dédié que l'utilisateur possède et qui est sécurisé grâce à un environnement d'exécution de confiance.

En outre, comme cela est couramment fait dans l'industrie et dans le milieu académique, nous supposons que les SGDP disposent d'une bonne connectivité et d'une aussi bonne disponibilité. De ce fait, les SGDP peuvent établir des connexions en pair-à-pair avec d'autres SGDP et peuvent être utilisés comme nœuds de calcul. Notre objectif dans ce papier est donc d'étudier des solutions reposant sur un réseau entièrement distribué de SGDP qui peuvent agir en qualité de nœud de calcul et/ou de contributeur et qui communiquent en pair-à-pair. Nous écartons volontairement les solutions qui nécessitent de recentraliser les données pendant un calcul car cela créerait dynamiquement des points de concentration qui seraient exposés aux mêmes risques que les serveurs centralisés.

Intégrer des environnements d'exécution de confiance augmente considérablement les protections contre des détenteurs de SGDP malveillants. Néanmoins, comme aucune mesure de sécurité ne peut être considérée comme inviolable, nous ne pouvons exclure l'hypothèse qu'une partie des SGDP du système soit corrompue, que ces SGDP collaborent, voire même qu'ils soient indétectables des autres SGDP honnêtes. De plus, puisque les calculs seront entièrement effectués par les SGDP, des fuites de données sensibles sont inévitables en la présence de nœuds corrompus — c'est-à-dire que certaines données pourraient être dévoilées dès lors qu'un nœud corrompu est sélectionné en tant que nœud de calcul.

Le but de ce papier est de mesurer la faisabilité de construire un système sécurisé, efficace et entièrement distribué de calcul, qui se repose sur un réseau de SGDP incluant des nœuds corrompus indétectables et qui collaborent. Pour y parvenir, nous proposons des mécanismes afin de réduire au maximum la fuite de données et faisons les contributions suivantes :

(1) Nous proposons une architecture en pair-à-pair de SGDP, appelée SEP2P (pour *Secure and Efficient P2P*), s'appuyant sur une Table de Hachage Distribuée et analysons les potentielles fuites de données personnelles. Nous montrons que (i) les opérations manipulant des données doivent être assignées de manière *aléatoire* et *vérifiable* — c'est-à-dire que l'assignation ne peut être influencée par un attaquant ; et que (ii) ces mêmes opérations doivent être

*atomiques* — c'est-à-dire réduites au maximum de sorte à minimiser la quantité de données sensibles manipulées.

(2) Nous nous concentrons sur le problème d'assignation aléatoire et vérifiable des opérations et proposons une solution générique (c'est-à-dire indépendante du calcul à effectuer) qui minimise le coût de vérification (par exemple, seulement 8 opérations cryptographiques asymmétriques dans un réseau contenant 1M de nœuds et dont 10K sont corrompus et collaborent).

(3) Nous évaluons expérimentalement la qualité et l'efficacité de nos protocoles. Plus particulièrement, nous montrons que notre protocole d'assignation aléatoire et vérifiable conduit à une fuite minimale, c'est-à-dire linéaire par rapport au nombre de nœuds corrompus, alors que le coût des mécanismes de sécurité reste très bas même avec un grand nombre de nœuds corrompus qui collaborent.

(4) Nous nous intéressons au sous-problème d'atomicité des opérations manipulant des données sensibles en fournissant trois ébauches de solutions pour les trois classes d'applications détaillées précédemment. Nous ne proposons pas de solution complète puisque ce problème dépend du cas d'usage considéré et nécessite donc d'être étudié, à chaque fois, en conséquence.

## RÉFÉRENCES


[1] Tristan Allard, Nicolas Anciaux, Luc Bouganim, Yanli Guo, Lionel Le Folgoc, Benjamin Nguyen, Philippe Pucheral, Indrajit Ray, Indrakshi Ray, and Shaoyi Yin. 2010. Secure personal data servers : a vision paper. *Proceedings of the VLDB Endowment* 3, 1-2 (2010), 25–35.

[2] Nicolas Anciaux, Philippe Bonnet, Luc Bouganim, Benjamin Nguyen, Philippe Pucheral, Iulian Sandu Popa, and Guillaume Scerri. 2019. Personal Data Management Systems : The security and functionality standpoint. *Information Systems* 80 (2019), 13 – 35.

[3] Nicolas Anciaux, Luc Bouganim, Philippe Pucheral, Yanli Guo, Lionel Le Folgoc, and Shaoyi Yin. 2014. MILo-DB : a personal, secure and portable database machine. *Distributed and Parallel Databases* 32, 1 (2014), 37–63.

[4] Blue Button. 2010. Find Your Health Data. Retrieved October 12, 2018 from https://www.healthit.gov/topic/health-it-initiatives/blue-button

[5] Cozy Cloud. 2013. Your digital home. Retrieved October 12, 2018 from https://cozy.io/en

[6] Yves-Alexandre de Montjoye, Erez Shmueli, Samuel S Wang, and Alex Sandy Pentland. 2014. openpds : Protecting the privacy of metadata through safeanswers. *PloS one* 9, 7 (2014), e98790.

[7] Fing. 2013. The mesinfos project explores the self data concept in france. Retrieved October 12, 2018 from http://mesinfos.fing.org/english

[8] Javier González, Michael Hölzl, Peter Riedl, Philippe Bonnet, and René Mayrhofer. 2014. A practical hardware-assisted approach to customize trusted boot for mobile devices. In *International Conference on Information Security*. Springer, 542–554.

[9] Anne-Marie Kermarrec and François Taïani. 2015. Want to scale in centralized systems ? Think P2P. *J. Internet Services and Applications* 6, 1 (2015), 16 :1–16 :12.

[10] Saliha Lallali, Nicolas Anciaux, Iulian Sandu Popa, and Philippe Pucheral. 2017. Supporting secure keyword search in the personal cloud. *Information Systems* 72 (2017), 1–26.

[11] MiData. 2011. The midata vision of consumer empowerment. Retrieved October 12, 2018 from https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment

[12] Nextcloud. 2016. Protecting your data. Retrieved October 12, 2018 from https://nextcloud.com

[13] European Parliament. 2016. General Data Protection Regulation. Law. Retrieved October 12, 2018 from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[14] Christian Priebe, Kapil Vaswani, and Manuel Costa. 2018. EnclaveDB : A Secure Database using SGX. In *EnclaveDB : A Secure Database using SGX*. IEEE, 0.

[15] Solid. 2018. Solid empowers users and organizations to separate their data from the applications that use it. Retrieved October 12, 2018 from https://solid.inrupt.com/

[16] Dai Hai Ton That, Iulian Sandu Popa, Karine Zeitouni, and Cristian Borcea. 2016. PAMPAS : Privacy-Aware Mobile Participatory Sensing Using Secure Probes. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management*. ACM, 4.

[17] Quoc-Cuong To, Benjamin Nguyen, and Philippe Pucheral. 2016. Private and scalable execution of SQL aggregates on a secure decentralized architecture. *ACM Transactions on Database Systems (TODS)* 41, 3 (2016), 16.

[18] J. C. Tomàs, B. Amann, N. Travers, and D. Vodislav. 2011. RoSeS : a continuous query processor for large-scale RSS filtering and aggregation. In *Proc. of the 20th ACM Conf. on Information and Knowledge Management*. 2549–2552.

# Streaming Saturation for Large RDF Graphs with RDFS Dynamic Schema Information

Mohammad Amin Farvardin
Université Paris-Dauphine, PSL, Research University
Paris, France
mohammad-amin.farvardin@lamsade.dauphine.fr

Dario Colazzo
Université Paris-Dauphine, PSL, Research University
Paris, France
dario.colazzo@lamsade.dauphine.fr

Khalid Belhajjame
Université Paris-Dauphine, PSL, Research University
Paris, France
khalid.belhajjame@lamsade.dauphine.fr

Carlo Sartiani
Università della Basilicata
Potenza, Italy
carlo.sartiani@unibas.it

## ABSTRACT

In the Big Data era, RDF data are produced in high volumes. While there exist proposals for reasoning over large RDF graphs using big data platforms, there is a dearth of solutions that do so in environments where RDF data are dynamic, and where new instance and schema triples can arrive at any time. In this work, we present the first solution for reasoning over large streams of RDF data using big data platforms. In doing so, we focus on the saturation operation, which seeks to infer implicit RDF triples given RDF Schema constraints. Indeed, unlike existing solutions which saturate RDF data in bulk, our solution carefully identifies the fragment of the existing (and already saturated) RDF dataset that needs to be considered given the fresh RDF statements delivered by the stream. Thereby, it performs the saturation in an incremental manner. Experimental analysis shows that our solution outperforms existing bulk-based saturation solutions.

## CCS CONCEPTS

• **Computing methodologies** → *Massively parallel algorithms.*

## KEYWORDS

RDF Saturation, RDF Streams, Big Data, Spark

# Scalable, Variable-Length Similarity Search in Data Series: The ULISSE Approach*

Michele Linardi
LIPADE, Paris Descartes University
michele.linardi@parisdescartes.fr

Themis Palpanas
LIPADE, Paris Descartes University
themis@mi.parisdescartes.fr

**ABSTRACT**

Data series similarity search is an important operation and at the core of several analysis tasks and applications related to data series collections. Despite the fact that data series indexes enable fast similarity search, all existing indexes can only answer queries of a single length (fixed at index construction time), which is a severe limitation. In this work, we propose ULISSE, the first data series index structure designed for answering similarity search queries of variable length. Our contribution is two-fold. First, we introduce a novel representation technique, which effectively and succinctly summarizes multiple sequences of different length (irrespective of Z-normalization). Based on the proposed index, we describe efficient algorithms for approximate and exact similarity search, combining disk-based index visits and in-memory sequential scans. We experimentally evaluate our approach using several synthetic and real datasets. The results show that ULISSE is several times (and up to orders of magnitude) more efficient in terms of both space and time cost, when compared to competing approaches.

# Reasoning about Disclosure in Data Integration in the Presence of Source Constraints (IJCAI'19)

**Michael Benedikt**[1] , **Pierre Bourhis**[2] , **Louis Jachiet**[2] and **Michaël Thomazo**[3]
[1]University of Oxford
[2]CNRS CRIStAL, Université Lille, Inria Lille
[3]Inria, DI ENS, ENS, CNRS, PSL University
{pierre.bourhis, louis.jachiet}@univ-lille.fr, michael.thomazo@inria.fr, michael.benedikt@cs.ox.ac.uk

## Abstract

Data integration systems allow users to access data sitting in multiple sources by means of queries over a global schema, related to the sources via map-pings. Datasources often contain sensitive informa-tion, and thus an analysis is needed to verify that a schema satisfies a privacy policy, given as a set of queries whose answers should not be accessible to users. Such an analysis should take into account not only knowledge that an attacker may have about the mappings, but also what they may know about the semantics of the sources. In this paper, we show that *source constraints* can have a dramatic impact on disclosure analysis. We study the problem of de-termining whether a given data integration system discloses a source query to an attacker in the pres-ence of constraints, providing both lower and upper bounds on source-aware disclosure analysis.

# Augmenting Analytic Datasets Using Natural and Aggregate-based Schema Complements

Rutian Liu
rutian.liu@sap.com
SAP France and Sorbonne Université, CNRS, LIP6
Paris

Eric Simon
eric.simon@sap.com
SAP France
Paris, France

Bernd Amann
bernd.amann@lip6.fr
Sorbonne Université, CNRS, LIP6
Paris, France

Stéphane Gançarski
stephane.gancarski@lip6.fr
Sorbonne Université, CNRS, LIP6
Paris, France

**ABSTRACT**

The production of analytic datasets is a significant big data trend and has gone well beyond the scope of traditional IT-governed dataset development. Analytic datasets are now created by data scientists and data analysts using big data frameworks and agile data preparation tools. However, despite the profusion of available datasets, it remains quite difficult for a data analyst to start from a dataset at hand and customize it with additional attributes coming from other existing datasets. This article describes a model and algorithms that exploit automatically extracted and user-defined semantic relationships for extending analytic datasets with new atomic or aggregated attribute values. Our framework is imple-mented as a REST service in the SAP HANA and includes a careful analysis and practical solutions for several complex data quality issues.

# Trustworthy Distributed Computations on Personal Data Using Trusted Execution Environments

Riad Ladjel
Inria, UVSQ, France
riad.ladjel@inria.fr

Nicolas Anciaux
Inria, UVSQ, France
nicolas.anciaux@inria.fr

Philippe Pucheral
Inria, UVSQ, France
philippe.pucheral@uvsq.fr

Guillaume Scerri
Inria, UVSQ, France
guillaume.scerri@uvsq.fr

***Abstract*—** Thanks to new regulations like GDPR, Personal Data Management Systems (PDMS) have become a reality. This decentralized way of managing personal data provides a de facto protection against massive attacks on central servers. But, when performing distributed computations, this raises the question of how to preserve individuals' trust on their PDMS? And how to guarantee the integrity of the final result? This paper proposes a secure computing framework capitalizing on the use of Trusted Execution Environments at the edge of the network to tackle these questions.

***Keywords*— *Data privacy; TEE; secure distributed computing***

# RDF graph anonymization robust to data linkage

## Anonymisation de graphes RDF résistante aux attaques par liens

Rémy Delanaux
Angela Bonifati
Romuald Thion
[prénom].[nom]@univ-lyon1.fr
Université Claude Bernard Lyon 1, LIRIS CNRS
Villeurbanne, France

Marie-Christine Rousset
[prénom].[nom]@imag.fr
Université Grenoble Alpes, CNRS, INRIA, Grenoble INP
Grenoble, France
Institut Universitaire de France
Paris, France

Depuis sa création, le paradigme du *Linked Open Data* (ou *LOD*, traduit par « web de données ouvertes » ou « données ouvertes liées ») a permis la publication de données ouvertes sur le Web et l'interconnexion de millions de ressources uniques, concrétisant un réel partage et un échange d'informations à une échelle globale. Le nuage de données ouvertes contenant toutes ces ressources, le *LOD cloud*, continue sa progression et contenait 1239 graphes RDF connectés par 16147 liens en mars 2019.

Depuis 2007, le nombre de ces graphes RDF publiés dans en tant que LOD a fortement augmenté, presque multiplié par un facteur 100. Toutefois, la participation de nombreuses organisations et institutions à ce mouvement d'ouverture est entravée par des problématiques de confidentialité des données sensibles ou personnelles. En effet, les données personnelles sont omniprésentes au sein de nombreux ensembles de données qu'elles possèdent, et les évolutions récentes dans la façon de traiter ces données et leur aspect critique (via notamment des règlements comme le RGPD européen ou d'autres réglementations liées aux données médicales ou gouvernementales) les rendent réticentes à publier leurs données.

Si des efforts ont déjà été fournis pour adapter des techniques d'anonymisation de données relationnelles au contexte du LOD [7, 12], comme des variations du modèle classique de $k$-anonymat [8, 10, 13], l'état de l'art et les travaux les plus récents se concentrent principalement sur des techniques de confidentialité différentielle pour les données relationnelles [4, 9].

Toutefois, la confidentialité différentielle n'est pas une solution très adéquate pour les données liées, se concentrant plus sur l'intégrité statistique des données anonymisées plutôt que sur des résultats précis et qualitatifs à des requêtes exécutées sur les données. Or ce dernier cas de figure représente le cas d'utilisation le plus fréquent des données liées interrogées via des *endpoints* SPARQL [1, 11]. Cette forme de confidentialité est donc utile quand des résultats agrégés liés à l'analyse des données peuvent être publiés, comme des statistiques sur des groupes de personnes. Si ceci peut avoir de nombreuses applications, il n'est pas suffisant dans le contexte de la publication de données respectueuse de la confidentialité (PPDP, *Privacy-Preserving Data Publishing*) [5] où la confidentialité des individus doit être protégée tout en assurant en parallèle que les données publiées seront utilisables en pratique.



Si les bases théoriques de la PPDP pour ses applications principales (comme l'anonymisation) ont déjà bien été étudiées pour les données relationelles (voir ce *survey* [5]), les bases théoriques de cette PPDP dans le contexte des données liées n'ont été évoquées que très récemment [6], en se concentrant principalement sur la complexité de calcul de l'opération visant à vérifier si les conditions permettant l'anonymisation de données liées étaient remplies. Deux problèmes fondamentaux y ont été étudiés, à savoir la *satisfaction d'une politique de confidentialité*, garantissant qu'un graphe isolé $G$ anonymisé ne révèle aucune information sensible, et le garantie de *sûreté face aux liens* qui assure que ce $G$ peut être publié avec des garanties prouvables face aux attaques par liens entre données.

Dans cet article, nous nous appuyons sur le cadre formel et théorique présenté en [6] en nous concentrant sur ce second besoin, c'est-à-dire la sûreté face aux attaques par lien, et nous présentons des algorithmes concrets calculant les opérations d'anonymisation qu'il nécessaire d'appliquer à n'importe quel graphe RDF pour obtenir cette garantie de sûreté quand ce graphe est lié à d'autres graphes externes du LOD cloud. En se basant sur la complexité de calcul de ce problème de sûreté ($AC_0$ en complexité de données sous l'hypothèse dite *open-world*), nous répondons au problème du calcul concret d'une séquence d'opérations d'anonymisation sûres qui mettent en place ces garanties de sûreté face aux attaques par liens. Dans cette étude, nous étudions plus particulièrement le cas des liens `sameAs` (c'est-à-dire des relations d'égalité entre ressources exprimées directement via des triplets RDF) qui peuvent être soit explicites dans le graphe original $G$, le liant à d'autres graphes externes, ou bien implicites et déduites via des mécanismes d'inférences exécutés sur $G$ lui-même.

Notre approche possède deux caractéristiques principales. En premier lieu, elle est *basée sur des requêtes* puisque les politiques de confidentialité ainsi que les opérations d'anonymisations sont spécifiées via respectivement des requêtes conjonctives et de requêtes de mise à jour SPARQL. Cela nous amène à définir une nouvelle définition de *sûreté*, légèrement différente de celle présentée en [6]. Dans un second temps, notre approche est *indépendante des données* car, pour une politique de confidentialité donnée (définie comme un ensemble de requêtes de confidentialité), nos algorithmes produisent des opérations d'anonymisation (sous la forme de requêtes SPARQL `UPDATE` pour la mise à jour de données et `DELETE` pour la suppression) avec la garantie que leur application à *n'importe quel graphe RDF* va satisfaire notre contrainte de sûreté. Ces propriétés sont extrêmement importantes pour la conception d'un cadre formel d'anonymisation de données liées, car elles nous permettent d'exploiter la maturité et les performances actuelles de moteurs

d'évaluation SPARQL pour évaluer ces requêtes et éditer le graphe à anonymiser, sans autre intermédiaire.

Nos contributions peuvent être résumées comme suit :

— nous établissons le problème de sûreté via les séquences concrètes d'opérations d'anonymisation nécessaire pour garantir cette sûreté ; nous fournissons donc une nouvelle définition de sûreté, *indépendante des données*, qui considère un ensemble de requêtes de confidentialité en entrée et non reliée aux instances de graphes concrètes. Cette définition se différencie donc de celle de base de sûreté face aux liens définie dans [6] ;

— nous fournissons des conditions suffisantes sous lesquelles une instance d'anonymisation est sûre étant donné un ensemble de requêtes de confidentialité ;

— nous concevons un algorithme d'anonymisation garantissant cette condition de sûreté, en étudiant sa complexité d'exécution. Nous montrons que cet algorithme s'exécute en temps polynomial en la taille (nombre de triplets total) de la politique de confidentialité donnée ;

— nous étudions les liens `sameAs` au sein de notre cadre d'étude, and montrons que notre algorithme est robuste face aux `sameAs` explicites ;

— nous présentons une étude expérimentale vouée à tester l'évaluation de nos algorithmes et la qualité des anonymisations fournies, confirmant leur efficacité et utilité en pratique.

Un des éléments majeurs exploité par notre approche est l'utilisation de *blank nodes*, qui sont l'équivalent de variables existentielles locales en RDF, et sont la clé de la prévention des attaques par lien. En effet, ces variables étant locales et donc uniques au graphe où elles sont créées, il n'est pas possible de lier à des données externes.

Nos algorithmes d'anonymisation fonctionnement en cherchant des termes critiques trouvés dans les corps des requêtes de confidentialité données, et en remplaçant les images de ces termes dans le graphe à anonymiser par des *blank nodes* tout en conservant les liens entre ces termes d'un triplet à l'autre. Car si notre objectif est de casser les jointures potentielles avec des données externes, il faut également maximiser l'utilité des données anonymisées. Pour cela, ces jointures internes qui ne sont pas dangereuses pour la confidentialité sont préservées. Cette approche est analogue à l'idée de « généralisation » développée dans les méthodes de $k$-anonymat. L'idée est de *neutraliser* une valeur sensible, tout en conservant l'intégrité des données et le plus d'information non-sensible qu'il est possible de garder. Cette approche nous permet notamment de conserver, en plus des garanties formelles de confidentialité, certains résultats d'utilité. En particulier, les requêtes de comptage associées aux requêtes de confidentialité données gardent une borne inférieure valide.

Notre approche peut être directement combinée avec d'autres approches de PPDP. Une fois le graphe RDF transformé via les opérations générées par nos algorithmes, il est tout à fait possible d'appliquer une autre anonymisation au graphe RDF obtenu. Il est par exemple possible de vérifier si le graphe RDF obtenu vérifie une certaine propriété de $k$-anonymat ou de confidentialité différentielle. L'adaptation plus directe d'approches dérivées du $k$-anonymat, notamment pour la généralisation plus pertinente de valeurs littérales (voir par exemple [12]), est prévue dans de prochains travaux. Elle peut également être combinée avec des techniques de réécriture de requêtes basées sur des ontologies pour des langages ontologiques du premier ordre comme RDFS [2] ou DL-Lite [3], en fournissant des requêtes de confidentialité réécrites en entrée de notre algorithme.

Différentes directions sont possibles pour les travaux futurs. La première est l'étude du risque de ré-identification inhérent à la délégation de la génération de nouveaux *blank nodes* au moteur d'évaluation SPARQL. Si ce choix est acceptable dans un cadre formel déclaratif comme le notre, il peut être vulnérable à une éventuelle attaque si le mécanisme de génération est connu par un adversaire externe. Nous prévoyons également d'étendre notre modèle de sûreté afin de gérer de la connaissance supplémentaire, par exemple le fait que certaines propriétés sont équivalentes ou qu'il existe une hiérarchie entre elles. Enfin, nous considérons l'étude de la version *dépendante des données* du problème de sûreté, et l'analyse de si ce problème amène ou non à des opérations d'anonymisation plus précises tout en garantissant la sûreté des données.

## ACKNOWLEDGMENTS

Ce projet a été soutenu par la région Auvergne-Rhône-Alpes via la programme ARC6 (« T.I.C. et Usages Informatiques Innovants ») qui a financé la thèse de doctorat de Rémy Delanaux ; par le LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) ; par le projet SIDES 3.0 (ANR-16-DUNE-0002) financé par le Programme Investissement d'Avenir (PIA) ; and par le programme Palse Impulsion 2016/31 (ANR-11-IDEX-0007-02) à l'Université de Lyon.

## RÉFÉRENCES

[1] Angela Bonifati, Wim Martens, and Thomas Timm. 2017. An Analytical Study of Large SPARQL Query Logs. *PVLDB* 11, 2 (2017), 149–161.

[2] Maxime Buron, François Goasdoué, Ioana Manolescu, and Marie-Laure Mugnier. 2019. Reformulation-based query answering for RDF graphs with RDF ontologies. In *ESWC, to appear*.

[3] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati. 2007. Tractable Reasoning and Efficient Query Answering in Description Logics : The *DL-Lite* Family. *J. Autom. Reasoning* 39, 3 (2007), 385–429.

[4] Cynthia Dwork. 2006. Differential Privacy. In *ICALP (2) (LNCS)*, Vol. 4052. Springer, 1–12.

[5] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing : A survey of recent developments. *ACM Comput. Surv.* 42, 4 (2010), 14 :1–14 :53.

[6] Bernardo Cuenca Grau and Egor V. Kostylev. 2016. Logical Foundations of Privacy-Preserving Publishing of Linked Data. In *AAAI*. AAAI Press, 943–949.

[7] Benjamin Heitmann, Felix Hermsen, and Stefan Decker. 2017. k–RDF-Neighbourhood Anonymity : Combining Structural and Attribute-based Anonymisation for Linked Data. In *PrivOn@ISWC*, Vol. 1951. CEUR-WS.org.

[8] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness : Privacy Beyond k-Anonymity and l-Diversity. In *ICDE*. IEEE Computer Society, 106–115.

[9] Ashwin Machanavajjhala, Xi He, and Michael Hay. 2016. Differential Privacy in the Wild : A tutorial on current practices & open challenges. *PVLDB* 9, 13 (2016), 1611–1614.

[10] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. *L*-diversity : Privacy beyond *k*-anonymity. *TKDD* 1, 1 (2007), 3.

[11] Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, and Adrian Bielefeldt. 2018. Getting the Most Out of Wikidata : Semantic Technology Usage in Wikipedia's Knowledge Graph. In *ISWC*. 376–394.

[12] Filip Radulovic, Raúl García-Castro, and Asunción Gómez-Pérez. 2015. Towards the Anonymisation of RDF Data. In *SEKE*. KSI Research Inc., 646–651.

[13] Latanya Sweeney. 2002. k-Anonymity : A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.

# Evaluation de la faisabilité de classification en utilisant les dépendances fonctionnelles

Marie Le Guilly
INSA Lyon, CNRS, LIRIS UMR5205
Villeurbanne, France
marie.le-guilly@insa-lyon.fr

Jean-Marc Petit
INSA Lyon, CNRS, LIRIS UMR5205
Villeurbanne, France
jean-marc.petit@insa-lyon.fr

Vasile-Marian Scuturici
INSA Lyon, CNRS, LIRIS UMR5205
Villeurbanne, France
marian.scuturici@insa-lyon.fr

## 1 CONTEXTE ET PROBLÉMATIQUE

Avec le nombre croissant d'outils et de librairies pour l'apprentissage automatique, il n'a jamais été aussi aisé d'utiliser des algorithmes de classification : quelques lignes de code suffisent pour appliquer des dizaines d'algorithmes différents, sur n'importe quel jeu de données. Il est ainsi "facile" pour les data scientists de produire des modèles d'apprentissage dans un temps limité. En contrepartie, les experts d'un domaine peuvent avoir l'impression que ces modèles sont des boîtes noires, qui pourraient marcher sur n'importe quelles données, sans vraiment comprendre pourquoi. Pour cette raison, et en lien avec l'interprétabilité des modèles d'apprentissage, il y a un réel besoin de réconcilier les experts avec les modèles produits, d'identifier le bon niveau d'abstraction, et des techniques pour les impliquer dans la construction du modèle.

Dans cet article, nous nous intéressons au problème de confiance dans les modèles d'apprentissage en utilisant les dépendances fonctionnelles. Nous défendons l'idée que les DFs caractérisent l'existence d'une fonction, qui est celle qu'un algorithme de classification cherche à définir. A partir de cette remarque, simple mais cruciale, nous proposons plusieurs contributions. Nous montrons d'abord comment les DFs donnent une borne supérieur pour l'exactitude (*accuracy*) d'un modèle de classification, en le confirmant via l'application de plusieurs algorithmes sur un ensemble de jeux de données de classification de UCI [1]. Nous montrons ensuite comment générer des jeux de données synthétiques "difficiles" pour la classification, afin de montrer que pour certains jeux de données, l'application de méthodes d'apprentissage n'a pas de sens. Enfin, nous proposons une solution pratique, passant à l'échelle, pour démontrer l'existence d'une fonction entre les attributs et la classe dans un jeu de données. Des expérimentations mettent en place cette solution sur des jeux de données réels appartenant à des entreprises.

## 2 PROPOSITIONS

### 2.1 Apprentissage et DFs

Si la classification et les dépendances fonctionnelles sont deux notions complètement distinctes, elles se basent toutes les deux sur la notion mathématique de fonction. En effet, un algorithme de classification chercher à définir la fonction qui associe au mieux chaque ligne de données à sa classe, en se basant sur une mesure d'erreur donnée. Les DFS, quant à elles, ne sont satisfaites que s'il existe une fonction allant des attributs de la partie gauche vers ceux de la partie droite de la DF. Ainsi, la notion de DF peut être appliquée à un jeu de données de classification, pour vérifier l'existence d'une fonction entre les attributs et la classe, et donc la satisfaction de la DF $\{attributs\} \rightarrow classe$.

Bien entendu, il est plus compliqué de définir la fonction en elle-même que d'en vérifier l'existence : mais il est justement intéressant de noter que dans la plupart des problèmes d'apprentissage, l'existence est supposée sans être formellement vérifier. Ainsi, cette étape supplémentaire pourrait permettre d'identifier des jeux de données pour lesquels un travail supplémentaire est nécessaire avant de construire le modèle d'apprentissage. La DF est alors un moyen d'évaluer la faisabilité de la classification sur le jeu de données considéré, permettant à des experts du domaine de mieux appréhender leur données et les possibilités d'apprentissage.

Pour la plupart des jeux de données, la DF ne sera pas entièrement satisfaite : en effet, il existera toujours des contre-exemples, les données réelles comportant nécessairement une part de bruit. Toute la question est alors d'estimer la proportion de contre-exemples dans le jeu de données. Lorsque celle-ci est trop importante, il faut alors étudier les données plus en détail, car cela affecte directement les performances des algorithmes d'apprentissage. Plusieurs mesures ont été définies pour mesurer la proportion de contre-exemple à une DF, notamment dans [2]. Parmi ces mesures, une est particulièrement intéressante, à savoir $G_3$, qui est définie de la manière suivante pour une relation $r$ et la DF $X \rightarrow Y$ :

$$G_3(X \rightarrow Y, r) = \frac{max(\{|s||s \subseteq r, s \models X \rightarrow Y\})}{|r|}$$

Cette mesure permet d'estimer la taille de la plus grande sous-relation, contenue dans la relation initiale, qui ne contienne aucun contre-exemple à la DF considérée. Il peut être démontré que pour un jeu d'apprentissage, en considérant la DF $\{attributs\} \rightarrow classes$, cette mesure est directement une borne supérieure pour l'exactitude de tout modèle de classification sur ce jeu de données (nombre de prédictions correctes sur le nombre total de prédictions). Ainsi, l'utilisation de la mesure $G_3$ en amont de la construction d'un modèle de classification permet directement d'estimer les performances maximales d'un tel modèle. Si la mesure est trop basse, un travail sur les données doit alors être envisagé, et l'étude détaillée des contre-exemples, c'est-à-dire des données limitant la satisfaction de la DF, peut être nécessaire.

Des expérimentations ont été menées en lien avec cette mesure $G_3$, qui a été évaluée pour plusieurs jeux de données issus de la base de données de UCI. En comparant les résultats obtenus aux mesures d'exactitude obtenues dans [1], il est apparu que les meilleures

1. archive.ics.uci.edu

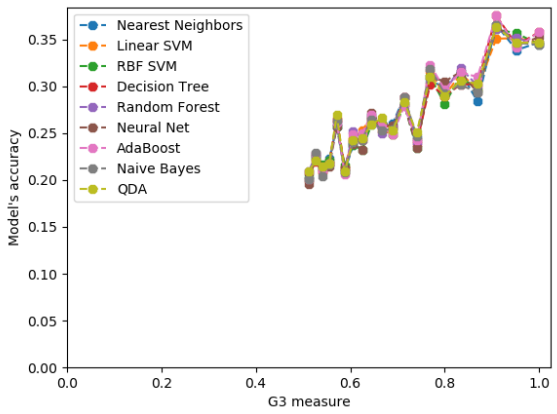


**FIGURE 1: Evolution de l'exactitude de modèle de classification en fonction de la mesure de $G_3$ sur des jeux de données synthétiques**

performances obtenues sont toujours très proches ou égales à la valeur de $G_3$ pour le jeu de données considéré.

## 2.2 Jeux de données difficiles

En se basant sur les observations faîtes sur la mesure $G_3$, il est apparu qu'il est également possible de générer des jeux de données ayant une valeur de $G_3$ aussi basse que l'on veut, et pour lesquels les modèles de classification ne pourraient donc jamais obtenir de performances satisfaisantes. De tels jeux permettent alors de mener des expérimentations supplémentaires, pour montrer le lien direct entre les performances des modèles de classification et $G_3$. De plus, ces jeux permettent d'obtenir des benchmarks pour les modèles de classification, qui peuvent être confrontés à des jeux difficiles : l'objectif est alors de produire des modèles dont l'exactitude est aussi proche de la mesure de $G_3$ que possible pour le jeu considéré.

La génération de tels jeux de données, pour lesquels la valeur de $G_3$ est contrôlée, se base sur l'introduction volontaire de contre-exemples. Ainsi, en partant d'un jeu de données n'en contenant pas, des réplications successives des lignes sont effectuées, en modifiant la classe associée à une ligne : comme cette classe constitue la partie droite de la dépendance fonctionnelle, répliquer la ligne en la changeant introduit nécessairement un contre-exemple à la DF $\{attributs\} \rightarrow classe$. Le nombre de réplication et de classes dans le jeu de données permet alors d'influer sur la valeur de $G_3$ de manière contrôlée.

Des expérimentations ont été menées pour étudier l'influence des différents paramètres pour la génération de ces jeux synthétiques, ainsi que sur l'impact de tels jeux pour les performances des algorithmes de classification. Ainsi, la figure 1 montre l'évolution de l'exactitude d'une dizaine d'algorithmes de classification (issus de la librairie scikit-learn [3]), en fonction de la mesure $G_3$ imposée pour la génération des jeux de données synthétiques : les résultats montrent clairement la manière dont $G_3$ limite les performances des différents modèles.

## 2.3 Jeux de données réels

L'approche proposée dans cet article peut se généraliser à n'importe quel jeu de données de classification : mais dans certaines situations pratiques, des limitations peuvent apparaître, pour lesquels nous avons alors proposé une solution. Tout d'abord, pour les jeux de données contenant beaucoup de valeurs continues, il est apparu que les dépendances fonctionnelles avaient beaucoup plus de chance d'être satisfaites : deux valeurs ayant beaucoup moins de chance d'être égale, au vu de l'immense nombre de valeurs possibles. Pour autant, du point de vue d'un expert métier, il peut être acceptable de considérer deux valeurs proches comme étant égale, au vu par exemple de l'incertitude de mesure : il est alors nécessaire de relâcher l'égalité stricte considérée pour les DF, afin d'identifier certains contre-exemples ayant du sens vis-à-vis de la réalité du terrain.

De plus, ce sont justement les contre-exemples qui ont une réelle valeur ajoutée pour les experts métiers : en effet, lorsque la valeur de $G_3$ est trop faible, ce sont les contre-exemples qui peuvent expliquer certains points de blocages, et qui peuvent permettre de proposer de nouvelles solutions pour améliorer la qualité des données pour l'apprentissage. Or, pour obtenir tous les contre-exemple, il est nécessaire de comparer les $n$ lignes du jeu de données deux par deux, ce qui représente $n * (n-1)/2$ comparaison. Dès que la taille du jeu de données est trop importante, ce calcul peut alors être rédhibitoire.

Dans cet article, nous proposons une solution hybride, permettant de répondre aux deux problématiques susmentionnées. Cette solution consiste à transformer le jeu de données initial, pour modifier les valeurs continues en valeurs discrètes, tout en réduisant le nombre total de lignes. Pour cela, dans un premier temps, chaque colonne est discrétisée, afin de regrouper les valeurs similaires entre elles : chaque valeur continue est alors remplacée par celle du groupe auquel elle appartient désormais : ainsi, deux valeurs continues proches se retrouveront dans le même groupe, et seront donc considérées comme égale une fois transformées par la discrétisation. Suite à cette première étape, le nombre possible de ligne différentes a diminué, et certaines lignes se retrouvent alors plusieurs fois dans le jeu de données. Il est alors possible de ne conserver qu'une seule fois cette ligne dans le jeu de données, en rajoutant une colonne "compteur", qui indique le nombre d'occurrences de cette ligne. Ainsi, le nombre de ligne diminue, rendant le nombre de comparaisons nécessaire aux calcul des contre-exemples moindre. Le compteur permet cependant de conserver les proportions initiales correctes, et donc de calculer la valeur $G_3$ en accord avec la distribution initiale des valeurs.

Cette approche permet donc de s'adapter à des jeux de données de taille conséquente, et d'adresser la spécificité des valeurs continues pour les dépendances fonctionnelles.

## RÉFÉRENCES

[1] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems ? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[2] Jyrki Kivinen and Heikki Mannila. 1995. Approximate inference of functional dependencies from relations. *Theoretical Computer Science* 149, 1 (1995), 129–149.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

# A relational framework for inconsistency-aware query answering

Ousmane Issa
University Clermont Auvergne
Clermont-Ferrand, France
ousmane.issa@uca.fr

Angela Bonifati
Lyon 1 University
Villeurbanne, France
angela.bonifati@univ-lyon1.fr

Farouk Toumani
University Clermont Auvergne
Clermont-Ferrand, France
farouk.toumani@uca.fr

## ABSTRACT

We introduce a novel framework for encoding inconsistency into relational tuples and tackling query answering for union of conjunctive queries (UCQs) with respect to a set of denial constraints (DCs). We define a notion of inconsistent tuple with respect to a set of DCs and define four measures of inconsistency degree of an answer tuple of a query. Two of these measures revolve around the minimal number of inconsistent tuples necessary to compute the answer tuples of a UCQ, whereas the other two rely on the maximum number of inconsistent tuples under set- and bag-semantics, respectively. In order to compute these measures of inconsistency degree, we leverage two models of provenance semiring, namely why-provenance and provenance polynomials, which can be computed in polynomial time in the size of the relational instances for UCQs. Hence, these measures of inconsistency degree are also computable in polynomial time in data complexity. We also investigate top-k and bounded query answering by ranking the answer tuples by their inconsistency degrees. We explore both a full materialized approach and a semi-materialized approach for the computation of top-k and bounded query results.

## KEYWORDS

Inconsistency, Why-provenance, Query Answering, Conjunctive Queries, Denial Constraints, Ranked Enumeration

**ACM Reference Format:**

Ousmane Issa, Angela Bonifati, and Farouk Toumani. 2019. A relational framework for inconsistency-aware query answering. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 12 pages.


*BDA'19, October 15–18, 2019, Lyon, France*

ACM ISBN 978-1-4503-9999-9/18/06…$15.00

# RDFPartSuite : Intégration du Partitionnement Logique lors du traitement de données RDF*

Jorge Galicia
LIAS/ISAE-ENSMA
Chasseneuil-du-Poitou, France
jorge.galicia@ensma.fr

Amin Mesmoudi
LIAS/Université de Poitiers
Poitiers, France
amin.mesmoudi@univ-poitiers.fr

Ladjel Bellatreche
LIAS/ISAE-ENSMA
Chasseneuil-du-Poitou, France
bellatreche@ensma.fr

## 1 INTRODUCTION

Resource Description Framework (RDF) est devenu une norme très populaire pour la représentation de données sous forme de graphes. Initialement, le modèle RDF a été conçu pour représenter l'information sur le Web et ce en utilisant seulement les triplets *<sujet,prédicat,objet>* comme structure de données basique. Par ailleurs, la flexibilité de RDF a motivé son utilisation dans d'autres domaines (par exemple la génétique) et aujourd'hui des bases de données RDF à grande échelle ont émergé. Dans cette optique, la recherche sur les systèmes de gestion de données RDF supportant le passage à l'échelle, distribués et parallèles, a pris de l'ampleur. La plupart de ces systèmes appliquent des algorithmes de partitionnement basés sur la notion de triplet comme une unité de distribution. Cette stratégie, purement physique, implique i) la perte de la structure du graphe de données et cause une dégradation des performances et ii) l'utilisation d'une stratégie de partitionnement difficilement personnalisable. Nous croyons que le fait de rassembler les triplets en fragments qui regroupent les mêmes entités logiques contribue non seulement à éviter l'exploration de données non pertinentes, mais aussi à créer des partitions RDF avec une signification logique. De plus, elle favorise l'intégration de techniques de partitionnement proposées dans les systèmes existants puisque les données sont manipulées à un niveau logique.

L'utilisation de la notion de triplet comme structure de données offre au modèle RDF la flexibilité nécessaire pour faciliter l'intégration de nouvelles données. Toutefois, ne pas appliquer un schéma même lorsque les données peuvent en avoir un explicitement, vient avec le prix de l'éparpillement de données.

Les fichiers RDF bruts stockent beaucoup de triplets (parfois des milliards) dans lesquels les faits de la même entité de haut niveau peuvent être éparpillé à travers le jeu de données. Cela pénalise non seulement le traitement des requêtes au sein des systèmes centralisés, mais aussi le choix d'une stratégie de partitionnement optimale dans les systèmes distribués. Comme il a été mentionné précédemment, les stratégies de partitionnement utilisées par plusieurs systèmes prennent les triplets comme unités de distribution et appliquent, par exemple, une fonction de hachage afin de distribuer les triplets, utilisés comme granularité la plus fine. Néanmoins, ces stratégies ne permettent pas de conserver la structure de graphe et affectent considérablement les performances. De plus, les stratégies dépendent de la façon dont les données sont stockées physiquement sur le disque (par exemple table de triplets [2], modèle de graphe [3]).

Le modèle relationnel offre des outils pratiques et puissants (e.g. superviseurs, langages) qui se reposent sur des structures logiques de haut niveau (relations). Ces outils offrent un confort au concepteur de base de données lorsqu'il partitionne les données horizontalement ou verticalement, peu importe la façon dont elles sont persistées physiquement. A long terme, nous visons à fournir des outils équivalents pour le partitionnement de données RDF.

Dans ce travail, nous donnons la définition formelle et détaillons les algorithmes nécessaires pour créer les entités logiques, que nous appelons fragments de graphe ($\mathcal{G}f$), utilisées comme unité de distribution pour les jeux de données RDF. Les entités logiques proposées sont harmonisées avec les techniques classiques de partitionnement par instances (horizontale) et par attributs (verticale) dans le modèle relationnel. Nous proposons des stratégies d'allocation de ces fragments, en considérant le cas où la réplication est possible. Nous discutons également la façon d'intégrer notre langage déclaratif de définition de partitionnement aux systèmes considérés comme l'état de l'art dans la gestion de données RDF à grande échelle. Nos expérimentations, sur des jeux de données synthétiques et réelles, montrent que les fragments de graphe évitent certains goulots d'étranglement lors du traitement de données RDF. Toutes nos techniques sont intégrées dans le même cadre applicatif, que nous avons appelé *RDFPartSuite*.

Nos contributions dans ce papier peuvent être résumées comme suite :

(1) La formalisation d'une dimension logique généralisant les partitions RDF par l'identification des entités de haut niveau.
(2) Une analyse des méthodes d'allocation intégrant à la fois le partitionnement par attributs et par instances dans les systèmes RDF.

## 2 APERÇU DE RDFPARTSUITE

Nous proposons de rassembler les données de la même manière que dans le modèle relationnel, en regroupant d'abord les tuples dans des entités de niveau supérieur. Ensuite, ces entités sont partitionnées au niveau de l'instance ou de l'attribut. Toutefois, dans RDF, il n'y a pas d'étape de conception indiquant des entités de haut niveau comme dans les bases de données relationnelles. Notre proposition permet de détecter des entités logiques implicites de haut niveau dans les données à partir de sources de données existantes basées sur le concept d'ensembles de caractéristiques [1]. L'ensemble de triplets, regroupant les triplets avec leurs arcs sortants et entrants dont chaque groupe partage le même ensemble de caractéristique, est nommé respectivement *fragments de graphes avant* $\overrightarrow{\mathcal{G}f}$ et *fragments de graphes arrière* $\overleftarrow{\mathcal{G}F}$.

*La version complète de ce document a été publiée dans les actes de la 21ème conférence internationale en Big Data Analytics and Knowledge Discovery DaWaK 2019.

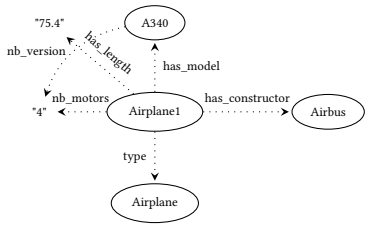


**Figure 1: Example du graphe RDF $G$**

De plus, le regroupement des triplets par les instances ou par les attributs contribue à l'amélioration des performances lors de l'exécution des requêtes. En effet, les partitions horizontales et verticales dans le modèle relationnel permettent de récupérer seulement des sous-ensembles de données en évitant l'exploration complètes des tables. De même, la construction de fragments de triplets évite de traverser le graphe entier en cherchant des correspondances spécifiques à une requête . Ces fragments deviennent les fragments d'allocation pour les stratégies décrites ci-après. De plus, nous décrivons un langage de définition qui simplifie la création de partitions en cachant des détails d'implémentation à l'utilisateur final.

## 2.1 Regroupement par instances

Nous observons qu'un ensemble de triplets avec le même sujet appartiennent à la même entité de haut niveau si les triplets entre ensembles partagent le même ensemble de prédicats. En général, les *fragments de graphes avant* peuvent être identifiés de façon unique par l'ensemble de ses arcs sortants (c.-à-d. prédicats). Ce concept est assez similaire aux ensembles de caractéristiques mentionnés par Neumann et al. [1].

Un fragment de graphes avant, représenté comme $\overrightarrow{\mathcal{G}f}$, rassemble des triplets de la même entité de haut niveau dans la base. Le graphe $G$, montré dans la Figure1, peut être représenté par 2 fragments de graphes avant regroupant les triplets avec `Airplane1` et `A340` comme sujet. Si $G$ avait plus de triplets rassemblés par le sujet dont l'ensemble caractéristique est le même que n'importe lequel des ensembles précédents, les triplets seraient ajoutés au même fragment de graphes avant.

## 2.2 Regroupement par attributs

Néanmoins, une organisation de données dans des fragments de graphes avant n'est pas optimale lors de la résolution des requêtes avec un nombre très réduit de prédicats. Le problème est tout à fait similaire à celui qui a motivé le partitionnement verticale dans les bases de données relationnelles auquel seuls les attributs pertinents d'une table sont accessibles. Cela nous a inspiré pour proposer un autre modèle d'organisation pour les données RDF que nous nommons *fragment de graphes arrière* $\overleftarrow{\mathcal{G}f}$ .

Les partitions verticales pour les données RDF sont obtenues en regroupant d'abord les triplets par leurs arcs entrants. En d'autres termes, nous groupons les triplets par leur objet. Dans la base de la Figure 1 nous obtenons 5 groupes `"4"`, `"75.4"`, `A340`, `Airplane` et `Airbus`. Comme aucun d'entre eux ne partage le même ensemble de caractéristiques, ces 5 ensembles ne peuvent plus être fusionnés.

## 2.3 Stratégies d'allocation

L'allocation des fragments est une étape obligatoire dans les systèmes distribués. L'utilisation d'une stratégie simple comme un round-robin, par exemple, peut ne pas produire une performance optimale. Tout d'abord parce que la taille des segments n'est pas uniforme, les données peuvent donc être inégalement réparties, ce qui entraîne des goulots d'étranglement sur certains sites. Deuxièmement, les requêtes impliquant plus d'un fragment peuvent être affectées par le coût d'envoi des résultats intermédiaires sur le réseau. Les coûts de réseau constituent le goulot d'étranglement des systèmes distribués et devraient être réduits au minimum pour améliorer les performances.

Nous avons formalisé le problème d'allocation et proposé une heuristique de partitionnement de graphe pour trouver une solution raisonnable.

## 2.4 Langage Déclaratif

La création de fragments (en avant et arrière) avant le partitionnement d'un fichier RDF brut contribue à intégrer une dimension logique au processus de partitionnement purement physique. Un langage déclaratif pourrait alors être utilisé pour décrire les processus de création, d'allocation et de partitionnement des deux types de fragments. Les déclarations sont faites de la même manière que les déclarations des tables avec le DDL (Data Definition Language) de SQL. Un exemple d'utilisation de ce langage déclaratif est le suivant:

```
CREATE {FORWARD,BACKWARD,BOTH}_FRAGMENTS
FROM [raw rdf file path]
WITH [max_size]
```

# 3 CONCLUSION

Dans cet article, nous introduisons une dimension logique au processus de partitionnement des graphes RDF. Nous formalisons et donnons un aperçu des algorithmes utilisés pour créer et allouer les entités logiques que nous avons nommés fragments de graphe $\mathcal{G}f$. Nous proposons un langage déclaratif pour faciliter la séparation de la couche de conception de la distribution de données de la couche d'exécution en cachant les détails de la mise en œuvre à l'utilisateur final. Nos expérimentations ont confirmé que la création de fragments logiques n'est pas plus coûteuse que la création de partitions avec des méthodes de partitionnement purement physiques. Nous avons utilisé des ensembles de données synthétiques et réelles pour montrer que les fragments de graphe permettent d'éviter certains goulots d'étranglement. De plus, nous avons montré l'intérêt des deux types de fragments de graphes (avant et arrière) lors de l'exécution des requêtes.

Nos travaux en cours incluent la prise en compte des requêtes dans la déclaration des partitions et la prise en compte des ensembles de données dynamiques (mises à jour).

# REFERENCES

[1] Thomas Neumann and Guido Moerkotte. 2011. Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins. In *Proceedings of the 27th ICDE, April 11-16, Hannover, Germany*. 984–994.
[2] Thomas Neumann and Gerhard Weikum. 2008. RDF-3X: a RISC-style engine for RDF. *PVLDB* 1, 1 (2008), 647–659.
[3] Lei Zou, Jinghui Mo, Lei Chen, M. Tamer Özsu, and Dongyan Zhao. 2011. gStore: Answering SPARQL Queries via Subgraph Matching. *PVLDB* 4, 8 (2011), 482–493.

# Assisted Classification Through Image- and Text-Based Event Detection (published in WEBIST'19)

Gabriela Bosetti
Előd Egyed-Zsigmond
Lucas Okumura-Ono
gabriela.bosetti@insa-lyon.fr
elod.egyed-zsigmond@insa-lyon.fr
lucas.okumura--ono@insa-lyon.fr
Université de Lyon. UMR 5205 CNRS
Villeurbanne, France

## ABSTRACT

Today, there are plenty of tools and techniques to perform text- or image-based classification of large datasets, targeting different lev-els of user expertise and abstraction. Specialists usually collaborate in projects by creating ground truth datasets and do not always have deep knowledge in Information Retrieval. This article presents a full platform for assisted binary classification of very large textual and text and image composed documents. Our goal is to enable human users to classify collections of several hundred thousand documents in an assisted way, within a humanly acceptable number of clicks. We propose a graphical user interface, based on several classification assistants: text- and image-based event detection, Ac-tive Learning (AL), search engine and rich visual metaphors to visualize the results. We also propose a novel query strategy in the context of Active Learning, considering the top unlabeled bi-grams and duplicated (e.g. re-tweeted) content in the target corpus to clas-sify. These contributions are supported not only by a tool whose code is freely accessible but also by an evaluation of the impact of using the aforementioned methods on the number of clicks needed to reach a stable level of accuracy.

## CCS CONCEPTS

• **Gestion de données spatiales, temporelles, scientifiques, mul-timédia**; • **Visualisation des données, exploration et interaction**; • **Réseaux sociaux, systèmes de recommandation et graphes de données**;

## KEYWORDS

information retrieval, assisted document classification, active learning, human-computer interaction

**ACM Reference Format:**
Gabriela Bosetti, Előd Egyed-Zsigmond, and Lucas Okumura-Ono. 2019.
Assisted Classification Through Image- and Text-Based Event Detection
(published in WEBIST'19). In Proceedings of WEBIST '19. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*WEBIST '19, September 18–20, 2019, Vienna, Austria*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

# Efficient Execution of Scientific Workflows in the Cloud through Adaptive Caching

Gaëtan Heidsieck
Inria & LIRMM
Univ. Montpellier, France
gaetan.heidsieck@inria.fr

Daniel de Oliveira
Institute of Computing, UFF
Rio de Janeiro, Brazil
danielcmo@ic.uff.br

Esther Pacitti
Inria & LIRMM
Univ. Montpellier, France
Esther.Pacitti@lirmm.fr

Christophe Pradal
CIRAD & AGAP
Univ. Montpellier, France
christophe.pradal@inria.fr

François Tardieu
INRA & LEPSE
Montpellier, France
francois.tardieu@inra.fr

Patrick Valduriez
Inria & LIRMM
Univ. Montpellier, France
Patrick.Valduriez@inria.fr

**ABSTRACT**

Many scientific experiments are now carried on using scientific workflows, which are becoming more and more data-intensive and complex. We consider the efficient execution of such workflows in the cloud. Since it is common for workflow users to reuse other workflows or data generated by other workflows, a promising ap-proach for efficient workflow execution is to cache intermediate data and exploit it to avoid task re-execution. In this paper, we propose an adaptive caching solution for data-intensive workflows in the cloud. Our solution is based on a new scientific workflow management architecture that automatically manages the storage and reuse of intermediate data and adapts to the variations in task execution times and output data size. We evaluated our solution by implementing it in the OpenAlea system and performing extensive experiments on real data with a data-intensive application in plant phenotyping. The results show that adaptive caching can yield ma-jor performance gains, *e.g.*, up to a factor of 4.5 with 6 workflow re-executions.

**KEYWORDS**

Adaptive Caching, Scientific Workflow, Cloud, Workflow Execution

# Clustering par modèle de mélange de Dirichlet : distribution et passage à l'échelle

Khadidja Meguelati
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
khadidja.meguelati@inria.fr

Benedicte Fontez
MISTEA, Montpellier SupAgro, Univ. Montpellier
Montpellier, France
benedicte.fontez@supagro.fr

Nadine Hilgert
MISTEA, INRA, Univ. Montpellier
Montpellier, France
nadine.hilgert@inra.fr

Florent Masseglia
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
florent.masseglia@inria.fr

## ABSTRACT

Le clustering avec des résultats précis est devenu un sujet d'intérêt majeur. Les modèles de mélange suivant un processus de Dirichlet, ou Dirichlet Process Mixture models (DPM) sont utilisés pour le clustering avec l'avantage de découvrir automatiquement le nom-bre de clusters et d'offrir de bonnes propriétés comme, par exemple, la convergence potentielle vers les clusters réels dans les données. Cependant, ces avantages se traduisent par des temps de réponse prohibitifs, ce qui nuit à son adoption et rend inefficaces les ap-proches centralisées pour le mettre en œuvre. Nous proposons DC-DPM[1], une solution de clustering parallèle qui s'adapte à des millions d'enregistrements tout en restant compatible avec le DPM, ce qui est le principal défi quand il s'agit de distribuer ce proces-sus. Nos expérimentations, tant sur des données synthétiques que réelles, illustrent les très bonnes performances de notre approche sur des millions d'enregistrements. L'algorithme centralisé, en re-vanche, ne passe pas à l'échelle et a trouvé sa limite dès 100000 enregistrements, où il a besoin de plus de 7 heures de calculs alors que notre approche prend moins de 30 secondes.

## KEYWORDS

Dirichlet Process Mixture Model, Clustering, Parallelism

## ACKNOWLEDGEMENTS

The research leading to these results has received funds from the Eu-ropean Union's Horizon 2020 Framework Programme for Research and Innovation, under grant agreement No. 732051.

[1] Khadidja Meguelati, Bénédicte Fontez, Nadine Hilgert, and Florent Masseglia. 2019. Dirichlet Process Mixture Models made Scalable and Effective by means of Massive Distribution. In *SAC 2019 - 34th Symposium On Applied Computing*. ACM/ SIGAPP, Limassol, Cyprus, 502–509. https://doi.org/10.1145/3297280.3297327

# Best Answers over Incomplete Data : Complexity and First-Order Rewritings

**Amélie Gheerbrant** and **Cristina Sirangelo**
Université de Paris, IRIF, CNRS, F-75013 Paris, France
{amelie, cristina}@irif.fr

## Abstract

Answering queries over incomplete data is ubiqui-tous in data management and in many AI applica-tions that use query rewriting to take advantage of relational database technology. In these scenarios one lacks full information on the data but queries still need to be answered with certainty. The cer-tainty aspect often makes query answering unfeasi-ble except for restricted classes, such as unions of conjunctive queries. In addition often there are no, or very few, certain answers, thus expensive com-putation is in vain. Therefore we study a relax-ation of certain answers called best answers. They are defined as those answers for which there is no better one (that is, no answer true in more possi-ble worlds). When certain answers exist the two notions coincide. We compare different ways of casting query answering as a decision problem and characterise its complexity for first-order queries, showing significant differences in the behaviour of best and certain answers. We then restrict attention to best answers for unions of conjunctive queries and produce a practical algorithm for finding them based on query rewriting techniques.

# Privacy-Preserving Informed Task Design in Crowdsourcing Processes

Joris Duguépéroux
Univ Rennes, CNRS, IRISA
Rennes, France
joris.dugueperoux@irisa.fr

Tristan Allard
Univ Rennes, CNRS, IRISA
Rennes, France
tristan.allard@irisa.fr

Antonin Voyez
Univ Rennes, CNRS, IRISA
Rennes, France
antonin.voyez@irisa.fr

**ABSTRACT**

Specialized worker profiles of high-skills crowdsourcing platforms may contain a large amount of identifying and possibly sensitive personal information (e.g., personal preferences, skills, available slots, available devices). Despite the recent interest in privacy-preserving crowdsourcing platforms - and more especially on privacy-preserving task assignment solutions - letting requesters design tasks that fit the skills and interests of workers while still providing sound privacy guarantees remains an open problem. In this paper, we combine together homomorphic encryption and differential privacy for computing an approximate distribution of the skills of a population of workers while satisfying differential privacy. The resulting distribution can be used by requesters in order to design tasks that fit as much as possible the underlying workers. In a nut-shell, we propose the PKD algorithm, we prove formally its security against *honest-but-curious* attackers, we analyze its complexity, and we demonstrate its quality and efficiency through an extensive experimental evaluation

# Aide à la Prise de Décision par Classement et Regroupement de Règles d'Association : *cas d'étude des clients* TOTAL

Idir Benouaret, Sihem Amer-Yahia
firstname.lastname@univ-grenoble-alpes.fr
CNRS, Univ. Grenoble Alpes

Senjuti Basu Roy
senjutib@njit.edu
New Jersey Institute of Technology
Neward, NJ, USA

Christiane Kamdem-Kengne, Jalil Chagraoui
firstname.lastname@total.com
TOTAL

## 1 INTRODUCTION

L'extraction de règles d'association [1] est l'une des techniques les plus populaires pour analyser le comportement d'achat des clients et obtenir des informations exploitables pour permettre une prise de décision. Les responsables marketing et chefs de produits de TOTAL mènent régulièrement des études sur les préférences des clients et les habitudes d'achat. Leur objectif est l'aide à la prise de deux décisions principales : *quels produits regrouper dans une offre promotionnelle et quels client cibler.* Cependant, lorsque le volume de données est très grand, cela peut conduire à une explosion des règles d'association; il faut donc utiliser des mesures de classement pour évaluer leur pertinence, tels que *Confiance*, *Piatetsky-Shapiro*, et *Lift*. Comme il existe de nombreuses mesures de classement (environ 35) [2, 4], il est nécessaire de déterminer quelle mesure de classement doit être utilisée pour quel type de tâche de prise de décision. Nous proposons de regrouper ces mesures de classement sur la base de leur similarité. Afin d'aligner le résultat de la fouille de règles d'association avec les besoins des praticiens, notre workflow consiste en quatre étapes : **Étape 1:** Nous donnons aux experts métier la possibilité d'exprimer et d'analyser les règles d'association d'intérêt. **Étape 2:** Nous regroupons les mesures de classement en 5 clusters synthétiques. **Étape 3:** Nous permettons aux experts du domaine de fournir un feedback sur ces clusters. **Étape 4:** À travers une étude utilisateur, nous discutons comment ce processus peut fournir des informations exploitables et permettre une aide à la décision.

## 2 EXTRACTION ET ÉVALUATION DES RÈGLES D'ASSOCIATION

Notre dataset représente les clients achetant des produits dans différentes stations-service qui sont géographiquement distribuées en France, pour une période de deux ans (de Janvier 2017 à Décembre 2018). Le dataset $\mathcal{D}$ est un ensemble d'enregistrements de la forme $\langle t, c, p \rangle$, ou $t$ est l'identtifiant de la transaction, $c$ est un client, et $p$ est un produit. L'ensemble de tous les identifiants de transactions est dénoté $T$. le dataset contient plus de 30 millions de transactions, couvrant 35 millions d'enregistrement et générées dans 3, 463 stations. Le ratio 30/35 est du au fait que, contrairement à d'autres domaines comme la grande distribution, la plupart des clients n'achètent que du carburant, et certains d'entre eux achètent des produits supplémentaires tels que le lavage ou les boissons. L'ensemble des clients $\mathcal{C}$ contient plus d'un million d'entrées. Chaque client est associé avec des attributs démographiques. Dans cette étude, nous nous focalisons sur 3 attributs : *age, sexe* et *localisation*. l'attribut *age* prend les valeurs $\{< 35, 35 - 49, 50 - 65, > 65\}$ et l'attribut *localisation* admet les régions françaises comme valeurs. L'ensemble des produits $\mathcal{P}$ contient plus de 37, 556 entrées. Chaque produit est associé à une catégorie de produit, telle que *carburant, lubrifiant* et *boissons*.

Afin de comprendre les habitudes d'achat des clients et leur proposer des offres pertinentes, les experts chez TOTAL souhaitent étudier deux types de pattern d'achats : ceux associant un ensemble de produits avec un produit cible (`prod_assoc`) et ceux associant les segments clients et une catégorie de produits (`demo_assoc`). Dans les deux cas, l'analyste spécifie une cible d'étude $\mathcal{B}$ et obtient des règles $\mathcal{A} \rightarrow \mathcal{B}$. La génération de règles d'association nécessite d'extraire d'abord des éléments fréquents de $\mathcal{T}$. Pour les extraire, nous utilisons *j*LCM [5], notre algorithme de fouille parallèle, distribué et open-source qui tourne sur MapReduce. La fouille des données se fait en deux étapes. Nous scannons le dataset $\mathcal{T}$ et construisons un dataset filtré, limité aux transactions contenant la cible $\mathcal{B}$. Ensuite, nous exécutons *j*LCM sur le dataset filtré. *j*LCM est récursif et extrait les ensembles de produits fréquents ainsi que leur fréquence. Les résultats de cette extraction contient le support de $\mathcal{B}$ et $\mathcal{A} \cup \mathcal{B}$ pour toutes les règles d'association qui nous intéressent. À ce stade, nous devons calculer le support de chaque ensemble de produits antécédent $\mathcal{A}$. Ainsi, dans une étape de post-traitement, nous scannons $\mathcal{T}$ pour calculer le support de tous les antécédents $\mathcal{A}$. Pour évaluer l'intérêt d'une règle d'association $\mathcal{A} \rightarrow \mathcal{B}$, plusieurs mesures d'intérêt telles que *Confiance, Lift, Support, Rappel, etc* qui servent à différentes analyses ont été proposées dans la littérature [2, 3]. Le Tableau 1 illustre un classement des top-5 règles pour la catégorie *Lubrifiants* et les top-5 pour le produit *Coca Cola* , classé suivant 3 mesures d'intérêt différentes.

## 3 CLASSEMENT ET CLUSTERING

Notre objectif est d'aider les analystes à sélectionner les règles d'association les plus exploitables, celles qui peuvent être utilisées pour promouvoir des produits ou cibler des clients spécifiques. Nous présentons une évaluation empirique des 35 mesures qui sont utilisées pour classer les règles d'association. Notre objectif principal est d'étudier la similitude des classements fournis par ces mesures d'intérêt, ce qui nous permet de les regrouper et de les rendre exploitable en réduisant leur nombre.

**Table 1: Top-5 des règles d'association pour `demo_assoc` et `prod_assoc`, selon les différentes mesures d'intérêt. Pour `demo_assoc`, les règles sont dénotées {age, sexe, région} → *catégorie cible*. Pour `prod_assoc`, {ensemble de produits} → *produit cible***

| by confidence | | by Piatetsky-Shapiro [6] | | by Recall | |
|---|---|---|---|---|---|
| {*50-65*, *M*,*Ile-de-France*} | → *Lubrifiants* | {*50-65*, *M*, ∗} | → *Lubrifiants* | {∗, *M*, ∗} | → *Lubrifiants* |
| {*50-65*, ∗, *Ile-de-France*} | → *Lubrifiants* | {∗, *M*,*Ile-de-France*} | → *Lubrifiants* | {*50-65*, ∗, ∗} | → *Lubrifiants* |
| {> 65, *M*,*Ile-de-France* } | → *Lubrifiants* | {∗, ∗,*Ile-de-France* } | → *Lubrifiants* | {*35-49*, ∗, ∗} | → *Lubrifiants* |
| {> 65, *M*, ∗} | → *Lubrifiants* | {∗, *F*, ∗} | → *Lubrifiants* | {*50-65*, ∗, ∗} | → *Lubrifiants* |
| {*50-65* , *M*,*Hauts-de-France* } | → *Lubrifiants* | {∗, *M*, ∗} | → *Lubrifiants* | {*35-49*, *M*, ∗} | → *Lubrifiants* |
| { *Bbq chips, Ham sandwish* } | → *Coca Cola* | {*Coffee* } | → *Coca Cola* | {*Coffee* } | → *Coca Cola* |
| {*Cheese sandwish, Bbq chips* } | → *Coca Cola* | {*Fuze Peche* } | → *Coca Cola* | {*Fuze Peche*} | → *Coca Cola* |
| {*Bbq chips, Salted chips* } | → *Coca Cola* | {*Insulated bottle*} | → *Coca Cola* | {*Mars legend* } | → *Coca Cola* |
| {*Chicken sandwish, Salted chips* } | → *Coca Cola* | {*Mars legend* } | → *Coca Cola* | {*Insulated bottle*} | → *Coca Cola* |
| {*Chicken sandwish, Bbq chips* } | → *Coca Cola* | { *Snickers* } | → *Coca Cola* | {*Snickers*} | → *Coca Cola* |

Nous nous appuyons sur les méthodes utilisées dans [4] afin de comparer des listes de règles d'association qui sont produites par différentes mesures d'intérêt. Les trois premières méthodes sont connues : *coefficient de corrélation de Spearman*, *le $\tau$ de Kendall* et Overlap@$k$. La dernière méthode *NDCC*, *Normalized Discounted Correlation Coefficient* est une mesure de corrélation proposée dans [4] inspirée de *NDCG, Normalized Discounted Cumulative Gain* et définie pour apporter plus d'importance aux résultats se situant en haut de liste.

Nous réalisons une analyse comparative des 35 mesures d'intérêt appliquées à nos deux scénarios. Nous générons un ensemble de règles d'association $A \longrightarrow B$, ou $B$ est un produit parmi 228 produits représentatifs qui ont été sélectionnés par nos analystes. Globalement, nous obtenons 253334 règles d'association. Nous calculons un classement par cible et par mesure d'intérêt. Nous calculons un matrice de corrélation de tous les classements en fonction de chacune des mesures de corrélation et nous calculons la moyenne sur tous les produits cibles. Ce qui nous donne une corrélation entre chaque paire de mesure d'intérêt. Nous utilisons ensuite un algorithme de regroupement hiérarchique avec lien moyen pour obtenir un dendrogramme des mesures afin d'analyser leurs similarités. Nous identifions 5 groupes de mesures similaires. $G_1$ contenant 18 mesures (dont *Lift, Confiance, Added value*) qui produisent des classements très similaires. Un deuxième groupe $G_2$ contenant 3 mesures (*Accuracy, Gini index, Least contradiction*) est similaire à $G_1$ selon le $\tau$ de Kendall. Mais cette similarité entre $G_1$ et $G_2$ est plus grande selon *NDCC*, ce qui signifie que les règles d'association en haut des classements $G_1$ et $G_2$ sont très similaires. Un troisième groupe $G3$ contenant 7 mesures (dont (*J-measure*) apparait, ainsi qu'un quatrième groupe $G_4$ contenant 5 mesures (dont *Piatetsky-Shapiro*), qui est similaire à $G_3$ selon *NDCC*. Enfin, nous obtenons un cinquième groupe $G_5$ contenant seulement les deux mesures *Recall* et *Collective strength*.

## 4 ÉTUDE

La réduction du nombre de mesures de classement nous permet de lancer une étude avec des experts du domaine chez `TOTAL`. Le but de l'étude est d'évaluer la capacité des mesures d'intérêt à classer les règles d'association en fonction des besoins des analystes. Plus précisément, nous souhaitons identifier les mesures d'intérêt les plus appréciées par nos experts. Chaque analyste choisit un scénario parmi `prod_assoc` ou `demo_assoc` pour lequel un produit ou une catégorie cible doit être choisi, respectivement. L'analyste reçoit alors des listes de règles d'association. Ni le nom de la mesure ni ses valeurs calculées ne sont révélés car nous voulions que les analyste évaluent les classements sans savoir comment ils ont été générés. Ils ont effectué 20 évaluations comparatives montrant deux classements à comparer avec à chaque fois les 10 premières règles d'association. Dans chaque cas, nous avons posé une question globale sur quel classement ils préféraient, et nous avons aussi demandé de marquer les règles pertinentes qui leur paraissent exploitables d'un point de vue marketing. En résumé, nous avons obtenu les feedbacks suivants : les classements qui favorisent la *Confiance* sont meilleurs pour déterminer quels produits regrouper lors d'une promotion, et les classements qui favorisent le *Rappel (Recall)* sont plus adaptés au cas où un produit est donné et le but est de trouver quels sont les meilleurs clients à cibler. Dans le cas de `prod_assoc`, le groupe le plus préféré était $G_1$, et une grande proportion de règles dans ce groupe étaient marquées pertinentes. Le prochain plus préféré dans ce même scénario est $G_2$. $G_1$ et $G_2$ favorisent tous les deux la *Confiance*, c'est-à-dire $P(B|A)$, et reflètent le cas où un produit est donné et l'objectif est de trouver les autres produits $A$ avec lesquels il faudrait les regrouper dans une promotion. Dans le cas de `demo_assoc`, le groupe le plus préféré était $G_5$, et une grande proportion de règles dans ce groupe étaient marquées pertinentes. Le second groupe le plus préféré dans ce même scénario est $G_4$. Les deux groupes $G_4$ et $G_5$ favorisent *Recall*, c'est-à-dire $P(A|B)$, et reflètent le cas où une catégorie de produit est donnée et le but est de trouver les clients qu'il faut cibler.

En résumé, ce travail a permis une étude plus approfondie du comportement des clients chez `TOTAL` et a donné les outils nécessaires pour le design de promotions de produits.

## REFERENCES


[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules between Sets of Items in Large Databases. In *Proc. SIGMOD*. 207–216.

[2] Liqiang Geng and Howard J. Hamilton. 2006. Interestingness Measures for Data Mining: A Survey. *ACM Comput. Surv.* 38, 3 (2006).

[3] Shin ichi Minato, Takeaki Uno, Koji Tsuda, Aika Terada, and Jun Sese. 2014. A Fast Method of Statistical Assessment for Combinatorial Hypotheses based on Frequent Itemset Enumeration. *Lect. Notes Artif. Int.* 8725 (2014), 422–436.

[4] Martin Kirchgessner, Vincent Leroy, Sihem Amer-Yahia, and Shashwat Mishra. [n.d.]. Testing Interestingness Measures in Practice: A Large-Scale Analysis of Buying Patterns. In *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016.*

[5] Martin Kirchgessner, Vincent Leroy, Alexandre Termier, Sihem Amer-Yahia, and Marie-Christine Rousset. [n.d.]. jLCM. https://github.com/slide-lig/jlcm. [Online; accessed 27-May-2016].

[6] Gregory Piatetsky-Shapiro. 1991. *Knowledge Discovery in Databases.* Menlo Park, CA: AAI/MIT.

# Patient trajectory prediction in the Mimic-III dataset, challenges and pitfalls

Jose F Rodrigues-Jr
junio@usp.br
University of Sao Paulo
Sao Carlos, SP, Sao Paulo

Gabriel Spadon
spadon@usp.br
University of Sao Paulo
Sao Carlos, SP, Sao Paulo

Bruno Brandoli
brunobrandoli@gmail.com
Dalhousie University
Halifax, Nova Scotia, Canada

Sihem Amer-Yahia
sihem.amer-yahia@imag.fr
Centre National de la Recherche Scientifique
Université Grenoble Alpes, Grenoble, France

## ABSTRACT

Automated medical prognosis has gained interest as artificial in-telligence evolves and the potential for computer-aided medicine becomes evident. Nevertheless, it is challenging to design an effec-tive system that, given a patient's medical history, is able to predict probable future conditions. Previous works, mostly carried out over private datasets, have tackled the problem by using artificial neu-ral network architectures that cannot deal with low-cardinality datasets, or by means of non-generalizable inference approaches. We introduce a Deep Learning architecture whose design results from an intensive experimental process. The final architecture is based on two parallel Minimal Gated Recurrent Unit networks working in bi-directional manner, which was extensively tested with the open-access Mimic-III dataset. Our results demonstrate significant improvements in automated medical prognosis, as mea-sured with Recall@k. We summarize our experience as a set of relevant insights for the design of Deep Learning architectures. Our work improves the performance of computer-aided medicine and can serve as a guide in designing artificial neural networks used in prediction tasks.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Ma-chine learning*; • **Applied computing** → **Health care informa-tion systems**; *Health informatics*.

## KEYWORDS

Neural Networks, Deep Learning, Patient Trajectory, Mimic-III

# Visualizing Health Tweets Over Regions and Timestamps

Bonpagna Kann
Sihem Amer-Yahia
Michael Ortega
Bonpagna.Kann@univ-grenoble-alpes.fr
Sihem.Amer-Yahia@univ-grenoble-alpes.fr
michael.ortega@imag.fr
CNRS, Univ. Grenoble Alpes

Jean-Louis Pépin
Sébastien Bailly
jpepin@chu-grenoble.fr
sbailly@chu-grenoble.fr
INSERM, Univ. Grenoble Alpes

## ABSTRACT

Social media has become one of the major data sources for studying our society. In healthcare, social media is thoroughly used to study people's discourse on particular ailments and derive insights on the impact of ailments on the patients' quality of life [6]. In this study, a total of nearly 800 million posts are retrieved for Twitter through pre-processing and running the Time-Aware Ailment Topic Aspect Model (T-ATAM) [8] with the purposes to predict diseases, symptoms and remedies for two chronic conditions: sleep apnea and chronic liver diseases. The study is conducted on tweets in English emitted during 2018 in European countries and the United States. The data has been processed using T-ATAM by regions, timestamps and treatment (Continuous Positive Airway Pressure, CPAP) in order to see the differences of the distributions of top diseases along with the top symptoms and remedies in different regions, timestamps, and before/during/after CPAP has been introduced. Based on approximately 331K tweets related to liver diseases and 1 million tweets on sleep apnea, we display various visualizations of statistics including world maps, word clouds and histograms. While depression and drinking are the leading symptoms for liver diseases, lack of night time sleep and overworking are considered as the main factors which contribute to sleep apnea.

## KEYWORDS

Social media, topic models, visualization, health data

# Improving Hamming distance-based fuzzy join in MapReduce using Bloom Filters

Thi-To-Quyen TRAN
Univ Rennes, CNRS, IRISA
Lannion, France
thi-to-quyen.tran@irisa.fr

Thuong-Cang PHAN
Cantho University
Cantho, Vietnam
ptcang@cit.ctu.edu.vn

Anne LAURENT
Univ Montpellier, LIRMM, CNRS
Monpellier, France
Anne.Laurent@lirmm.fr

Laurent d'Orazio
Univ Rennes, CNRS, IRISA
Lannion, France
laurent.dorazio@univ-rennes1.fr

## ABSTRACT

A fuzzy (or similarity join) combines all pairs of tuples for which the distance is lower than or equal to a prespecified threshold $\varepsilon$ from one or several relations. Fuzzy join has been studied by many researchers because of its practical application. However, join operation is quite costly and may even not be possible to compute on large scale databases. Besides that, MapReduce has become an increasingly popular massively parallel framework. In this paper, we thus propose the optimization for MapReduce algorithms to process fuzzy joins of binary strings using Hamming Distance. In particular we propose to use an extension of Bloom Filters to eliminate the redundant data, reduce the unnecessary comparisons, and avoid the duplicate output. We compare and evaluate analytically the algorithms with a cost model.

## KEYWORDS

Fuzzy join, Similarity join, MapReduce, Bloom Filter

## 1 INTRODUCTION

In recent years, researches has focused on the problem of efficient joins in large-scale parallel environments. The first results, concerning the equi-join [7], impose strong constraints on the data (one of the sets having to be small enough to be distributed to all the machines used for the treatment) or their organization (sorting according to the join attribute, placement of data on specific nodes), leading to many data transfers (some unnecessary) and heavy workload on machines or requiring multiple (expensive) execution phases. A fuzzy (or similarity join) arose in many applications [3–5, 9, 10]. When dealing with a very large amount of data, fuzzy join becomes a challenging problem in a distributed parallel computing environment with the expensive cost of data shuffle. As a result, the data redundancy is very difficult to accept. Vernica et al. [11] proposed a similarity join method using 3-stage MapReduce which utilized the prefix filtering method to support set-based similarity functions. Metwally et al. [6] proposed a 2-stage algorithm VSMART join for similarity join on set, multisets and vector. Afrati et al. [1] proposed multiple algorithms to perform fuzzy join in a single MapReduce stage. While recent studies on the fuzzy join have the common limitations as redundancy and duplication of data, the filter-based approaches in our recent studies [7, 8] can solve these problems. Our team was interested in using Bloom Filters (BF) [2] and Intersection Filter [8]. The idea is to filter irrelevant data as soon as possible to reduce data transfers and load on different machines. This study, therefore, focuses on a theoretical analysis of various Hamming distance-based similarity join algorithms in MapReduce, and their cost comparison in a map-reduce-shuffle computation.



## 2 BLOOM FILTER-BASED FUZZY JOINS

Previous algorithms [1] generate intermediate elements that may be not relevant to the join process in the map phase, because they do not match with any similar record in the input dataset. In our research, we propose to integrate BF into the join algorithms to improve performances. Our common solution consist of two stages: (1) Pre-processing stage builds a filter using join key value set, (2) Join processing stage uses this filter on a distributed cache to eliminate non-similar elements of the input dataset during the map phase before sending it to the reducers.

BF-BH1 Algorithms: During the map phase, $BH_1$ generates all elements within a distance $d$ from $s$ and send them to the reducers to combine them with similar input records. It is easy to see that not all elements in the $B_s(d)$ belong to $S$. Our approach integrates $BF(S)$ to remove elements in $B_s(d)$ that do not belong to $S$ before sending it to the reducers.

$$s \xrightarrow[BF(S)]{map} \begin{cases} (s,-1) \\ (t,s), \quad \forall t \in B_s(d) \cap S, t < s \end{cases}$$

With the assumption that hash operation performs in unit time, the pre-processing cost on all input records is $k|S|$. However, this cost can be amortized by streaming or caching techniques. Each membership test also uses $k$ hash functions, so the map cost for each record is $kB(d)$. If we note $\delta_S$ the ratio of similar records of $S$, $f_{BF(S)}$ the false positive probability of the BF of $S$, then the cost to transfer intermediate data from mappers to reducers is

$$D_{BF-BH1} = |S|[\delta_S B(d) + f_{BF(S)}(1-\delta_S)B(d)] < |S|B(d)$$

BF-Splitting algorithm: The Splitting algorithm generates redundant data by sending each record to $d+1$ reducers. In fact, each

**Table 1: Summary of costs for various Hamming distance-based join algorithms**

| Approche | Pre-processing | Map cost per element | # Reducers | Communication | Processing |
|---|---|---|---|---|---|
| Naive | 0 | $J \approx \sqrt{K}$ | $K$ | $\|S\|\sqrt{K}$ | $\|S\|^2$ |
| BH1 | 0 | $B(d)$ | $n = 2^b$ | $\|S\|B(d)$ | $\|S\|^2 B(d)/2^b$ |
| BF-BH1 | $k\|S\|$ | $kB(d)$ | $n = 2^b$ | $D_{BF-BH1} < \|S\|B(d)$ | $D_{BF-BH1}\|S\|/2^b$ |
| Splitting | 0 | $d+1$ | $(d+1)2^{b/(d+1)}$ | $(d+1)\|S\|$ | $(d+1)\|S\|^2/2^{b/(d+1)}$ |
| BF-Splitting | $k\|S\|$ | $kB(d)(d+1)$ | $(d+1)2^{b/(d+1)}$ | $D_{BF-Splitting} < (d+1)\|S\|$ | $D_{BF-Spitting}\|S\|/2^{b/(d+1)}$ |

record just need to be sent to some identified reducers if all its actual similar elements present in $S$ and its sub-strings are known. As a solution we propose to combine Ball Hashing, Splitting and BF.

In the join stage, the mapper generates all elements in the ball of radius $d$ around each input record $s$. By the membership test in $BF(S)$, it determines which elements $\{t \neq s\}$ in $B_s(d)$ may actually be similar to $s$. Then each of them is divided in to $d+1$ equal-length substrings $\{s_i\}$ and $\{t_i\}$, $i = 1..(d+1)$. For each $s_i$, if there exists a substring $t_i$ of $t$ in the intersection of $S$ and $B_s(d)$ that matches with $s_i$, the pair $(s_i, s)$ will be generated outputs, and then $t$ will never be considered again.

$$s \xrightarrow[BF(S)]{map} (s_i, s) \begin{cases} s_i \subset s \\ \forall t \in B_s(d) \cap S \\ \exists t_i \subset t \equiv s_i \end{cases}$$

The map cost is $k|S|B(d)(d+1)$. Each record is sent only if there is actual similar elements, with a small false positive. The communication cost is

$$D_{BF-Splitting} = [\delta_S|S| + f_{BF(S)}(1-\delta_S)|S|] < (d+1)|S|$$

The reducers collect, test the distance, and output records as in the Naive algorithm. However, in such an approach, each similar pair in $S$ is sent to at most one reducer, solving the duplicated output problem without a lexicography test. The total computation cost for reducers is $D_{BF-Splitting}|S|/2^{b/(d+1)}$ In order to compare the costs of different algorithms, it adapts a previous model $(M, C, R)$[1], where $M, R, C$ are used to measure the effectiveness of an algorithm.

Table 1 summarizes the costs of the different algorithms. According to the processing cost, Naive algorithm is the most expensive solution, but its cost is independent with the change of distance. With respect to the communication cost, Splitting algorithm is the best approach, while Ball Hashing is the most suitable solution to processing cost. However, Ball Hashing is sensitive to distance. With the greater the distance $d$, the number of elements in $B(d)$ increases dramatically. Integrating BF in the algorithms implies the following changes according the $(M, C, R)$ model:

- The pre-processing cost is incurred by reading the input to generate $BF(S)$. However this cost can be amortized, especially using streaming or caching techniques (e.g Spark).
- The map phase use $k$ hash functions for the membership test. In the BF-Splitting, the map phase generates $B_s(d)$ for each input record.
- The number of reducer does not change.
- Using $BF(S)$, redundant elements are eliminated, thus the communication cost is reduced. This also leads to a decrease of the computation cost on reducers.

As a conclusion, no algorithm is the best. Choosing a solution depends on the context. However, in a parallel and distributed environment, communication cost is one of the most important factors. Experiments in our previous studies [7, 8] have proved that filtering can significantly improve execution times.

## 3 CONCLUSIONS

In this paper, we study theoretical details for the fuzzy join algorithms based on Hamming distance measure in MapReduce, applied for $b$-bit strings input dataset. We propose the optimization for the Ball Hashing and Splitting algorithms, and show the comparison through the MapReduce cost model. Our approaches eliminate the redundant intermediate data, reduce the unnecessary comparisons and avoid the data duplication. For the fuzzy join of multiple input datasets, Intersection filter [8] is applied instead of Bloom filter. Our optimizations may be extended in the cache or streaming supported framework to reuse the preprocessing cost. Future work includes further validation of our works, comparison with other approaches and extension the research for other fuzzy join algorithms.

## REFERENCES


[1] Foto N. Afrati, Anish Das Sarma, David Menestrina, Aditya Parameswaran, and Jeffrey D. Ullman. 2012. Fuzzy Joins Using MapReduce. In *ICDE*. 498–509.
[2] Burton H. Bloom. 1970. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Commun. ACM* 13, 7 (1970), 422–426.
[3] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic Clustering of the Web. In *WWW*. 1157–1166.
[4] Monika Henzinger. 2006. Finding Near-duplicate Web Pages: A Large-scale Evaluation of Algorithms. In *SIGIR*. 284–291.
[5] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. 2007. Detectives: Detecting Coalition Hit Inflation Attacks in Advertising Networks Streams. In *WWW*. 241–250.
[6] Ahmed Metwally and Christos Faloutsos. 2012. V-SMART-Join: A Scalable MapReduce Framework for All-Pair Similarity Joins of Multisets and Vectors. *CoRR* abs/1204.6077 (2012). http://arxiv.org/abs/1204.6077
[7] Thuong-Cang Phan, Laurent d'Orazio, and Philippe Rigaux. 2016. A Theoretical and Experimental Comparison of Filter-Based Equijoins in MapReduce. *TLDKS* 25 (2016), 33–70.
[8] Thuong-Cang Phan, Laurent d'Orazio, and Philippe Rigaux. 2013. Toward Intersection Filter-based Optimization for Joins in MapReduce. In *Cloud-I*. 2:1–2:2.
[9] Mehran Sahami and Timothy D. Heilman. 2006. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In *WWW*. 377–386.
[10] Ellen Spertus, Mehran Sahami, and Orkut Buyukkokten. 2005. Evaluating similarity measures: A large-scale study in the Orkut social network. In *SIGKDD*. 678–684.
[11] Rares Vernica, Michael J. Carey, and Chen Li. 2010. Efficient Parallel Set-similarity Joins Using MapReduce. In *SIGMOD*. 495–506.

# A Unified Approach to Biclustering Based on Formal Concept Analysis

Nyoman Juniarta, Miguel Couceiro, Amedeo Napoli
Université de Lorraine, CNRS, Inria, LORIA
F-54000, Nancy, France
nyoman.juniarta@loria.fr,miguel.couceiro@inria.fr,amedeo.napoli@loria.fr

## ABSTRACT

In a matrix representing a numerical dataset, a bicluster is a subma-trix whose cells exhibits similar behavior. Biclustering is naturally related to Formal Concept Analysis (FCA) where concepts corre-sponds to maximal and closed biclusters in a binary dataset. In this paper, we propose a unified characterization of biclustering algo-rithms using FCA and pattern structures, an extension of FCA for dealing with numbers and other complex data. We show how several types of biclusters, e.g. constant-value, constant-column, constant-row, gradual patterns, etc. are related to partition pattern structures based either on object or attribute partitions. Moreover we discuss the potential and the generality of the present approach w.r.t. numer-ical approaches. Finally we present a series of experiments which show that the approach is well-founded and efficient.

# Range Query Processing for Monitoring Applications over Untrustworthy Clouds

Hoang Van Tran
Université de Rennes 1, IRISA
Lannion, France

Tristan Allard
Université de Rennes 1, IRISA
Rennes, France

Laurent D'Orazio
Université de Rennes 1, IRISA
Lannion, France

Amr El Abbadi
UC Santa Barbara
California, USA

## 1 INTRODUCTION

Privacy is a major concern in cloud computing since clouds are considered as untrusted environments. In this study, we address the problem of privacy-preserving range query processing on clouds. Over the last years, different approaches have attempted to strike a trade-off between security and practical efficiency [2, 5, 6]. Index-based schemes [3, 8, 9] have also been proposed to increase query performance while ensuring strong security. Nevertheless, prior schemes cannot cope with the high rate of incoming data. To this end, we propose an extension of the PINED-RQ [9] that enables the building of a *secure* index over sensitive data. The index is built at a trusted component (e.g., collector) before being published to the cloud. We choose PINED-RQ since it offers strong privacy protection and efficient range query processing, compared to its counterparts [3, 8]. Nonetheless, PINED-RQ has to publish data in batches and partially processes data at the collector. Consequently, bottlenecks may occur as incoming data arrives at a high rate. Therefore, in this paper, we propose PINED-RQ++, to mainly prevent potential bottlenecks at the collector. In particular, we reverse the process of constructing PINED-RQ's index, that allows to immediately send new data to the cloud without sacrificing privacy.

## 2 PINED-RQ++

We focus on the architecture as depicted in Figure 1. Data generators produce raw data and send them to a collector. The incoming data are then pre-processed prior to being sent to a cloud. A consumer poses range queries to the cloud. In PINED-RQ++, we assume that the cloud is *honest-but-curious* while the other components are trusted. Thus, the adversary can use all information exchanged between the cloud and the other trusted components to deduce anything in a computationally-feasible way.

At the beginning, an index template is built at the collector. Whenever a new tuple arrives, the index template is updated with that tuple. Next, the tuple is encrypted and forwarded to the cloud. When the index template is published at a later time, the cloud associates it with unindexed data to produce a secure index as described in [9].

A query is mainly processed at the cloud which holds both indexed and unindexed data at time. As a consumer issues a query, it is first evaluated on indexed data (as in PINED-RQ's query processing [9]), the result of this evaluation and all the un-indexed data are returned to the consumer. In parallel, the removed tuples overlapping the query range at the collector are also sent back to the consumer. Finally, the consumer decrypts and filters the returned data for the final results.

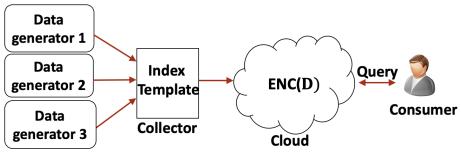


**Figure 1: Proposed architecture**

**Index template.** The structure of an index template is typically the same as in PINED-RQ [9]. However, since initially there are no real data, its count variables only contain Laplace noise [4] and its leaves have no pointers. Such count variables and pointers are updated during a publishing time interval, which is defined as the period from when an index template is initiated to when it is published.

**Matching table.** When an index template is published, the cloud associates it with unindexed data to form a PINED-RQ's index for those data. To prepare for this association, the collector needs to keep the pointers between unindexed data and leaves. To do that, a simple way is to mark the ciphertext of a new tuple by the id of the leaf node to which the tuple belongs, and send the marked ciphertext to the cloud. Later, the cloud can rebuild pointers from marked ciphertexts when the index template is published. However, these marked ciphertexts reveal the real pointers between unindexed data and leaves during a time interval. PINED-RQ++ consequently discloses more extra information, e.g., the actual distribution of the incoming time of real data, when compared to PINED-RQ.

To prevent the leakage of such information, we use unique random numbers which are viewed as temporary ids of tuples and a matching table (see Figure 2). The first column in this matching table stores leaves' id while each row of the second column holds the temporary id of tuples belonging to the corresponding leaf node. In particular, when a tuple arrives, the collector encrypts it, generates a unique random number, and sends the *<random number, ciphertext>* pair to the cloud. This number is stored in the corresponding row in the matching table at the collector. The randomness guarantees that no useful information about the index template is leaked

to the adversary. When an index template and its matching table are published, the cloud simply loops over the matching table and replaces random numbers with the leaves' pointers.

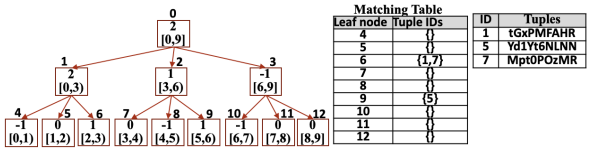


**Figure 2: Perturbed index template and matching table**

**Noise management.** One challenge to our approach is that the collector initiates and perturbs the index template without any existing data. This means that no tuples are available to be deleted in case of negative noise. Also, when leaves receive positive noise, the collector can initially generate dummy tuples. However, when those dummy tuples are published to ensure privacy protection must be considered. To this end, we present two approaches as follows.

Regarding positive noise, the collector sends dummy data according to the actual distribution of the sending time of the real tuples. With this approach, when a *<random number, ciphertext>* pair comes to the cloud, the adversary cannot distinguish which pair is dummy or real data.

For negative noise, if a leaf node initially receives negative noise *c*, the collector moves the first *c* tuples (when they arrive) of that leaf node to the corresponding overflow array. At publication time, the collector randomly fills all overflow arrays with dummy tuples and sends these overflow arrays to the cloud. Notably, the movement of tuples only occurs at the collector and the adversary does not know which nodes receive negative noise. Thus, the privacy of such movements is also preserved.

**Index template update management.** The main goal is to guarantee that the counts of PINED-RQ++'s index template are the same as those of PINED-RQ's index when the index template is published. In PINED-RQ [9], a leaf node's count represents the number of tuples falling within its interval. The count of internal nodes and the root is a summation of their children's counts. All counts are then perturbed by Laplace noise [4]. In contrast, an index template only contains noise at first and it will increase its counts as soon as a tuple arrives at the collector. Basically, when a tuple arrives at the collector, the leaf node to which the tuple belongs is determined. Then, the count of that leaf node and all its ancestors will be increased by 1.

## 2.1 Parallel PINED-RQ++

Since the collector, that is assumed to be a small private node with a few cores, updates the index template, its throughput would be impacted as the rate of incoming tuples increases. We therefore parallelize the construction of the index template to improve the collector's throughput. Instead of keeping only one index template for updating, we distribute that process on multiple *shared-nothing* machines (e.g., computing nodes).

## 2.2 Privacy Analysis

As compared to PINED-RQ [9], both PINED-RQ++ and its parallel version only leak extra information of *<random number, ciphertext>* pairs and the time, when such pairs arrive at the cloud, to the adversary. Random numbers will not disclose any information about the data. Besides, as discussed in Section 2, when a pair arrives at the cloud, an adversary cannot distinguish between a dummy and a real tuple. Thus, PINED-RQ++'s privacy protection is similar to that of PINED-RQ.

# 3 EVALUATION RESULTS

**Benchmark Environment.** We implemented our proposal on a real cluster and evaluated it on four metrics namely network traffic, the time needed to publish an index (template), time response latency, and throughput. We use two real datasets NASA log [1] and Gowalla [7] for our experiments.

**Results.** The maximum data rate is reduced by up to ~2.7x (NASA) and ~2.5x (Gowalla) in PINED-RQ++. Moreover, the publishing time is reduced by up to ~35x for NASA (514972 tuples) and ~16x reduction for Gowalla (1116907 tuples). For response latency, our approach performs slightly better with large ranges (100%) because PINED-RQ's collector experiences higher workload for processing consumers' queries. Regarding the parallel PINED-RQ++, the throughput at the collector is improved as we increase the number of computing nodes used. The 10-computing node setting gives the highest throughput, increasing by about 5.3x (to ~13K tuples/second) for NASA and about 4.4x (to ~66K tuples/second) for Gowalla when compared to the non-parallel version.

# 4 CONCLUSION

We developed PINED-RQ++ to address the challenges of data streams for processing range queries in clouds and the scalability problems of the prior schemes. The experimental results show that our solution provides better performance than PINED-RQ while privacy is also protected. Furthermore, we introduce the parallel PINED-QR++ that helps significantly enhance the throughput at the collector. In the future, we will consider on improving query performance, caching techniques and a dynamic adaptation to the query workload.

# REFERENCES

[1] 1996. NASA Log. http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html
[2] Rakesh Agrawal, Jerry Kiernan, Ramakrishnan Srikant, and Yirong Xu. 2004. Order Preserving Encryption for Numeric Data. In *SIGMOD*. 563–574.
[3] Ioannis Demertzis, Stavros Papadopoulos, Odysseas Papapetrou, Antonios Deligiannakis, and Minos Garofalakis. 2016. Practical Private Range Search Revisited. In *SIGMOD*. 185–198.
[4] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC*. 265–284.
[5] Bijit Hore, Sharad Mehrotra, Mustafa Canim, and Murat Kantarcioglu. 2012. Secure Multidimensional Range Queries over Outsourced Data. In *VLDB*. 333–358.
[6] Bijit Hore, Sharad Mehrotra, and Gene Tsudik. 2004. A Privacy-preserving Index for Range Queries. In *VLDB*. 720–731.
[7] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
[8] Rui Li, Alex X. Liu, Ann L. Wang, and Bezawada Bruhadeshwar. 2014. Fast Range Query Processing with Strong Privacy Protection for Cloud Computing. *In VLDB* (2014), 1953–1964.
[9] Cetin Sahin, Tristan Allard, Reza Akbarina, Amr El Abbadi, and Esther Pacitti. 2018. A Differentially Private Index for Range Query Processing in Clouds. In *ICDE*. 857–868.

# Generalization of Schema Mappings for Transformation Reuse

Paolo Atzeni
Università Roma Tre, Italy
atzeni@dia.uniroma3.it

Luigi Bellomarini
Banca d'Italia, Italy
luigibellomarini@bancaditalia.it

Paolo Papotti
EURECOM, France
papotti@eurecom.fr

Riccardo Torlone
Università Roma Tre, Italy
torlone@dia.uniroma3.it

## 1 SCHEMA MAPPING REUSE

Schema mappings are widely used as a tool for data exchange and integration. Although there are systems supporting users in the creation of mappings [2], designing them is still a time-consuming task. We notice that data transformation scenarios are often defined over schemas that are different in structure but similar in semantics. It follows that a great opportunity to reduce the effort of transformation design is to *reuse existing schema mappings*. Unfortunately, there is no obvious approach for this problem. Consider the following example.

*Example 1.1.* A central bank maintains a register with balance data from all national companies (Figure 1). This register has schema **G**, with relation `Balance` storing, information for each company. External providers send data to the bank in different forms. Provider A adopts schema $S_A$, with relation $R_A$ for companies (firms), whose code refers to relation `Activity`. Provider B adopts schema $S_B$, with a relation $R_B$ for companies (enterprises), whose code refers to relation `Location`. Data is moved from $S_A$ and $S_B$ into **G**, by using two schema mappings:

$\sigma_A$: $R_A(f, g, z, s)$, Activity$(s, d) \rightarrow$ Balance$(f, g, z, d)$.

$\sigma_B$: $R_B(e, g, s, c, a)$, Location$(a, n) \rightarrow$ Balance$(e, g, n, s)$.

The example shows a *data exchange scenario* where the differences in the mappings are due to the structural differences between $S_A$ and $S_B$, which are, on the other hand, semantically very similar. Moreover, every new data provider (e.g., $S_C$ in the figure) would require the manual design of a new, ad-hoc mapping, even if there is a clear analogy with the already defined mappings.

Our goal is to reuse $\sigma_A$ and $\sigma_B$ and avoid the definition of a new mapping for $S_C$. The intuition is to collect all available mappings in a repository; then, for any new pair of schemas (e.g., $S_C$ and **G** in the figure), query such repository to retrieve a suitable mapping. Unfortunately, this form of direct reuse is complicated by the nature of schema mappings. A mapping characterizes the constraint between a pair of schemas at a level of detail that enables both logical reasoning and efficient execution. Yet, a simple variation in a schema, such as a different relation or attribute name or a different number of attributes, makes it not applicable. Our experiments show that mappings from a corpus of 1.000 schema mappings can be reused for new pairs of schemas only in 20% of cases. To be reusable, a mapping should be described in a way that is independent of its specificities while harnessing the essence of the constraint so as to work for similar schemas.

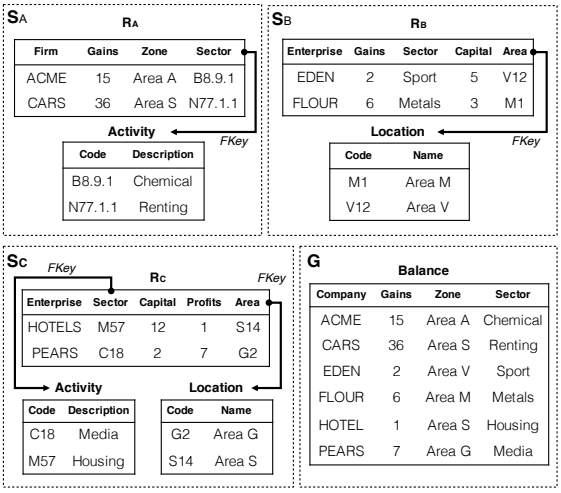


**Figure 1: A data transformation scenario**

*Example 1.2.* Consider a "generic" mapping $\Sigma_A$, obtained from $\sigma_A$ by replacing names of the relations and attributes with variables. It could be informally described as follows:

$\Sigma_A$: *for each* relation $r$ *with* key $f$ and attributes $g$, $a$, $s$
*for each* relation $r'$ *with* key $s$ and attribute $d$
with a foreign key constraint from $s$ of $r$ to $s$ of $r'$
*there exists* a relation $r''$ *with* key $f$ and attributes $g$, $a$, $d$.

If instantiated on $S_A$, the generic mapping $\Sigma_A$ expresses a mapping to **G** that is the same as $\sigma_A$. This solution seems a valid compromise between precision, i.e., the ability to express the semantics of the original mapping, and generality, as it can be applicable over different schemas. However, $\Sigma_A$ falls short of the latter requirement, as it is not applicable on $S_B$. Indeed, there are no constraints on attribute $g$ and $a$, and so they could be bound to any of `Gains`, `Sector` and `Capital`, incorrectly mapping `Capital` into the target.

*Example 1.3.* Consider a generic mapping using constants:

$\Sigma_B^H$: *for each* relation $r$ *with* key $e$ and attributes $g$, $s$, $c$, $a$
*for each* relation $r'$ *with* key $a$ and attribute $d$
with a foreign key constraint from $a$ of $r$ to $a$ of $r'$
*where* $g = Gains$, $s \neq Gains$, $s \neq Capital$, $c \neq Gains$, $c \neq Sector$
*there exists* a relation $r''$ *with* key $e$ and attributes $g$, $d$, $s$.

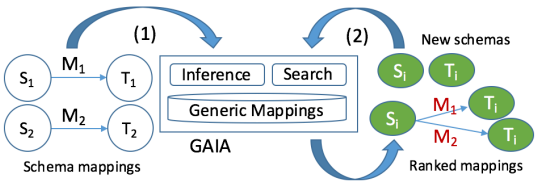


**Figure 2: The architecture of GAIA.**

This generic mapping is precise enough to correctly describe both $\sigma_A$ and $\sigma_B$ and can be re-used with other schemas.

The example shows a combination of attributes, identified by constraints on their names and role, that form a correct and useful generic mapping. Once pinpointed, generic mappings can be stored in a repository, so that it is possible to use them for a new scenario. In our example, given $S_C$ and G, the generic mapping $\Sigma_B^H$ can be retrieved from the repository and immediately applied.

There are three main challenges in this approach.

- We need a clear characterization of what it means for a generic mapping to correctly describe and capture the semantics of an original schema mapping.
- As a generic mapping is characterized by a combination of conditions on attribute names and roles, for a given schema mapping there is a combinatorial number of generic mappings. We need a mechanism to generate them.
- For a new scenario (e.g., new schemas), there is an overwhelming number of generic mappings that potentially apply, with different levels of "suitability". We need efficient tools to search through and choose among them.

## 2 A SYSTEM FOR MAPPING REUSE

We propose GAIA, a system for mapping reuse, which supports two tasks, as shown in Figure 2: (1) infer generic mappings, called *meta-mappings*, from input schema mappings and store them in a repository; (2) given a source and a target schema, return a ranked list of possible mappings between these schemas from the meta-mappings in the repository. GAIA provides the following contributions:

- The notion of *fitness*: a semantics to precisely characterize and check when a meta-mapping is suitable for a reuse scenario.
- An algorithm to *infer* meta-mappings from schema mappings; this algorithm is used to populate a repository of meta-mappings.
- An approach to reuse based on: (i) the search, in the repository of available meta-mappings, for those that *fit* a new pair of source and target schemas and (ii) the construction, from the retrieved meta-mappings, of possible mappings to be proposed to the user.

We recall next the notion of schema mapping and introduce that of *meta-mapping*. While the former notion models specific transformations, the latter introduces an abstraction over mappings [3, 4] and models *generic* mappings between schemas. Building on these notions, we illustrate the functionalities of our system. Because of space limitation, details are in the full version of the paper [1].

*Example 2.1.* A mapping between $S_A$ and G is the st-tgd $\sigma_A$ in the Introduction. The application of the chase to the instance of $S_A$ using $\sigma_A$ enforces this dependency by generating one tuple in the target for each pair of source tuples matching the LHS of the dependency. The result includes the first two tuples in relation `Balance` in Figure 1.

**Meta-mappings.** A *meta-mapping* describes generic mappings between relational schemas and is defined as a mapping over the catalog of a relational database [4]. Specifically, in a relational meta-mapping, source and target are both defined over the following schema, called *(relational) dictionary*: Rel(*name*), Att(*name*, *in*), Key(*name*, *in*), FKey(*name*, *in*, *refer*). An instance S of the dictionary is called *m-schema* and describes relations, attributes and constraints of a relational schema S. Given a source m-schema S and a meta-mapping $\mathscr{M}$, a target m-schema T is generated by applying the chase procedure to S using $\mathscr{M}$.

**From meta-mappings to mappings.** Given a source schema S and a meta-mapping $\Sigma$, it is possible not only to generate a target schema, but also to automatically obtain a schema mapping $\sigma$ that represents the *specialization* of $\Sigma$ for S and T [4]. The *schema to data exchange transformation* generates from S and $\Sigma$ a complete schema mapping made of S, a target schema T (obtained by chasing the m-schema of S with the meta-mapping), and an s-t tgd $\sigma$ between S and T. The correspondences between LHS and RHS of $\sigma$ are derived from the provenance information computed during the chase step.

Where to obtain useful meta-mappings to feed the repository? In GAIA those are obtained from existing schema mappings. Given an s-t mapping $\sigma$, the inference algorithm in GAIA generates all the meta-mappings that are fitting for the original scenario. This means that by applying the schema to data exchange transformation with the original source schema and the fitting meta-mapping, we obtain again the original mapping (up to renaming of variables).

*Example 2.2.* By chasing schema $S_A$ with meta-mapping $\Sigma_A$, we obtain a target m-schema with a placeholder name (denoted by labelled null $\perp_R$) and the following mapping from $S_A$ to $\perp_R$:

$$\sigma : R_A(f, g, z, s), \text{Activity}(s, d) \rightarrow \perp_R(f, g, z, d).$$

We get back, up to a renaming of the target relation, mapping $\sigma_A$, from which $\Sigma_A$ originates, i.e., it is a fitting meta-mapping.

**Re-using mappings.** Given a repository of meta-mappings, obtained from existing mappings, and a new pair of source and target schemas, GAIA supports mapping reuse by generating for the unseen schemas a ranked list of suitable mappings.

*Example 2.3.* For the running example, GAIA first generates fitting meta-mappings from $\sigma_A$ between $S_A$ and G. Once $S_B$ and G are given as input, the system scans the repository and identifies $\Sigma_A$ as a fitting meta-mapping from $S_B$ to G. This meta-mapping is then instantiated to generate a schema mapping between them.

Experiments show that our approach efficiently identifies useful mappings for a new scenario, while the conventional approach requires a user to completely define a new transformation.

## REFERENCES

[1] P. Atzeni, L. Bellomarini, P. Papotti, and R. Torlone. Meta-mappings for schema mapping reuse. *PVLDB*, 12(5):557–569, 2019.
[2] R. Fagin, L. M. Haas, M. A. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis. Clio: Schema mapping creation and data exchange. In *Conceptual Modeling*, 2009.
[3] M. A. Hernández, P. Papotti, and W. C. Tan. Data exchange with data-metadata translations. *PVLDB*, 1(1):260–273, 2008.
[4] P. Papotti and R. Torlone. Schema exchange: Generic mappings for transforming data and metadata. *Data Knowl. Eng.*, 68(7):665–682, 2009.

# 6 Résumés des articles doctorants

# Data sharing in presence of access control policies

Juba Agoun
juba.agoun@univ-lyon1.fr
Universite de Lyon, Universite de Lyon 1, CNRS LIRIS

## ABSTRACT

In the context of data analysis and data integration, very often information from different and autonomous sources are shared. Sources use their own schema and their own access control policies. We consider the case where data sources decide to share information by specifying entity matching rules[1] between their contents. A query to a given data source is rewritten to produce queries to other data sources that share information with that data source. This entity-matching oriented and policy-oriented rewriting preserves local data source policies.

## KEYWORDS

Data sharing, Data integration, Entity matching, Record matching, Access control, Query rewriting.

## 1 INTRODUCTION

A large volume of information is shared between several sources for data analysis purposes [7]. Data sharing is one of the configurations that allows sources using different schemas to share information. It uses data-level mappings which differs from data exchange and data integration [6]. In a data exchange setting (e.g., [2]), with a source schema $S$ and a target schema $T$, mappings captured by *source-to-target dependencies* are used to populate a target schema with the data of a source schema and specify how and what source data should appear in $T$. In data integration, the approach consists in defining a single entry point to sources by specifying a mapping between the global schema and each source schema. The mapping can be achieved by using one of the well-known approaches, namely, GAV (Global As View) or LAV (Local As View) (e.g., [4]).

Data in heterogeneous sources may overlap or may be closely associated to each other. Entity matching is used to identify the same real-world object from different sources even if it is represented differently in the different sources (different formats, spellings,...) or containing errors. The data heterogeneity problem arises when the same-real object is represented using different identifiers/features in different sources. For example, a patient may be uniquely identified using the social security number (*SSN*) in a hospital and a *donor_id* in a blood bank service.

Sharing data from different sources could potentially reveal sensitive information [3]. Indeed, each source enforces its own security policy to control the access to its content. Access controls may also differ from one source to another as they are independent and autonomous. However, some of the information that are shared could be sensitive in a given source and not sensitive in another one. For some pieces of data that may be closely associated and similar[2], the access could be denied in one source, whereas it may be guaranteed in another source. Hence, one faces a violation of a security policy.

There are several motivating scenarios for data sharing meeting with privacy in numerous real-world applications. For instance, sharing healthcare data could improve scientific research. It can, thus, enable early detection of disease outbreak [1]. However, obtaining consent to use piece of information can be prohibitive. Indeed, a disclosure of sensitive information can lead to a major damage to individuals.

## 2 EXAMPLE

Let us consider two institutions, a *hospital* and a *blood bank* that agree on sharing some subsets of their data.

The *hospital* stores the patient data in a relation we denote *patient*( *SSN, name_p, address_p, city_p, sex, blood_pressure, blood_glucose, height_weight, diagnosis* ) where *SSN* denotes the social security number, *name_p* denotes the name, *address_p* denotes the patient's address, *city_p* denotes city, *sex* denotes gender, *blood_pressure* is the measured blood pressure, *blood_glucose* is the blood glucose level, *height_weight* is the height and weight, and *diagnosis* is the type of the illness of the patient.

The *blood bank institution* stores in the relation *donor*( *id_d, donor _name, donor_address, donor_city, gender, blood_pressure, blood_glucose, h_w, number_donation*) information related to donors. The attribute *id_d* is the donor's id, *donor_name* is the donor's name, *donor_address* is the address, *donor_city* stands for the city, *gender* indicates the donor's gender, *blood_pressure* refers to a measured blood pressure, *blood_glucose* refers to the blood glucose level, *h_w* denotes the donor's height and weight, the *number_donation* represents the accumulated number of blood donations made by a donor.

The two relations *patient* and *donor* have an *attribute alignment* that maps the *patient* attributes *name_p, address_p, city_p, sex, blood_pressure, blood_glucose, height_weight* to *donor_name, donor_address, donor_city, gender, blood_pressure, blood_glucose, h_w* of the relation *donor*, respectively.

The two databases display entity heterogeneity as they do not share a common identifier (*see tables 1 and 2* ). Some information describing a donor might be associated to a given patient in the hospital database if both tuples refer to the same real-world individual. Indeed, a donor may previously visited the hospital for some medical treatment. For instance, in the given blood bank database instance, a donor's name may be stored as `John A. Smith`, while in the hospital, it is recorded as `Smith, John`. The two entities, patient and donor, represent the same real-world person with similarities

[1]Entity matching rules specify under which conditions two records from different sources are considered as a *match* and represent the same real-world object.

[2]Semantically equivalent, they denote the same real-world object

| SSN | name_p | address_p | city_p | sex | blood_pressure | blood_glucose | height_weight | diagnosis |
|---|---|---|---|---|---|---|---|---|
| 738-77-8987 | Bob Tracy | 3 rue emile zola | lyonn | M | 120/79 | 73 | 162/71 | headache |
| 358-87-9526 | Smith, John | 06 bis rue notre dame | paris 6 | M | 126/76 | 71 | 182/85 | stomachache |
| 852-37-9526 | Tim McCall | 43 av. des Postes | Lille center | M | 124/75 | 131 | 175/42 | diabetes |
| 436-44-0945 | Jeane Henri | 48, rue du Four | 75006 Paris | F | 146/97 | 69 | 156/52 | Hypertension |

**Table 1: Instance of *patient***

| id_d | donor_name | donor_address | donor_city | gender | blood_pressure | blood_glucose | h_w | number_donation |
|---|---|---|---|---|---|---|---|---|
| 455 | Robert Tracy | 03 rue emile Zola | lyon | Male | 120/79 | 71 | 162/72 | 2 |
| 589 | John A. Smith | 06 bis rue notre dame | paris | Male | 127/77 | 70 | 181/89 | 4 |
| 996 | Timothy McCall | 43 aveunue des Postes | lille | Male | 121/73 | 136 | 176/53 | 0 |
| 195 | Marine.P Jolio | 48 B.v pierre marion | marseille | Female | 116/77 | 72 | 163/55 | 1 |

**Table 2: Instance of *donor***

in name, address, city and sex. The similarities are computed by an entity matching rule $\phi_{EM}$. Given two tuples $p \in patient$ and $d \in donor$ the entity matching rule $\phi_{EM}$ is expressed as follows :

$$\begin{aligned}\Phi_{EM} = \; & p[name\_p] \approx_{(Jaro,78)} d[donor\_name] \wedge \\ & p[address\_p] \approx_{(Levenshtein,72)} d[donor\_address] \wedge \\ & p[city\_p] \approx_{(Smith-Waterman,77)} d[donor\_city] \wedge \\ & p[sex] \approx_{(jaro,70)} d[gender].\end{aligned}$$

The entity matching rule is computed over a subset of the aligned attributes where $\approx_{(f,\epsilon)}$ is the corresponding similarity function $f$ and $\epsilon$ is a threshold. The two records $p$ and $d$ *match* iff $\Phi_{EM}$ is evaluated to *true*.

Tables 1 and 2 show instances of the relations *patient* and *donor*, respectively. Please note that there are some similar records because they satisfy $\Phi_{EM}$. For instance, the record *Tim McCall* with the *SSN* "852-37-9526" in the *patient* instance has a similar record with the *id_d* "996" in the *donor* instance, as they satisfy $\Phi_{EM}$ with this evaluation:

$$\begin{gathered}Jaro(\text{ Tim McCall, Timothy McCall}) = 90 \wedge \\ Levenshtein(\text{ 43 av. des Postes, 43 aveunue des Postes}) = 76 \wedge \\ Smith - Waterman(\text{Lille center, lille}) = 80 \wedge \\ jaro(M, Male) = 75.\end{gathered}$$

Each data source exposes a security policy expressed as a set of access control rules. In our example, we consider the rules that express *forbidden views* [5], also called *secret views*. The hospital database denies access to the combination of SSN and diagnosis (*rule* $r_1$), and the combination of name and diagnosis (*rule* $r_2$) for all the patients. The rules $r_1$ and $r_2$ express the access control policy $\Pi_{patient}$ associated with the relation patient :

| | | |
|---|---|---|
| $r_1$ : | **Deny** | SSN, diagnosis |
| | **From** | patient |
| $r_2$ : | **Deny** | name_p, diagnosis |
| | **From** | patient |

In the other hand, the blood bank database denies access to the combination of name and the blood pressure (*rule* $r'_1$), and the association of the donor's id and the blood pressure (*rule* $r'_1$) for all the donors living in Lille. The access control policy $\Pi_{donor}$ associated with the relation donor is as follows:

| | | |
|---|---|---|
| $r'_1$ : | **Deny** | donor_name , blood_glucose |
| | **From** | donor |
| | **Where** | donor_city = "lille " |
| $r'_2$ : | **Deny** | id_d, blood_glucose |
| | **From** | donor |
| | **Where** | donor_city = "lille " |

Now, assume that we want to retrieve all the men's patient with their name, city, blood_pressure, blood_glucose and their height/weigh from *patient*. Such a query could be expressed as:

| | | |
|---|---|---|
| $q$ : | **Select** | name_p, city_p, blood_pressure, blood_glucose, height_weight |
| | **From** | patient |
| | **Where** | sex = "M" |

The attribute alignment and the entity matching rule between *patient* and *donor* might provide sufficient information to translate the query $q$ posed against the hospital database to a query $q'$ posed against the blood bank database. The derived query $q'$ could have the following form:

| | | |
|---|---|---|
| $q'$ : | **Select** | donor_name, donor_city, blood_pressure, blood_glucose, h_w |
| | **From** | donor |
| | **Where** | gender = "Male" |

The evaluation of $q$ discloses the patient's record "*Tim McCall*" whereas its match "*Timothy MacCall*" was hidden at the evaluation of the translated query $q'$. In general, such a translation requires a thorough analysis of the context and it should be based on a rigorous process to ensure the preservation of local access control policies. In our work, we are investigating a methodology, together with relevant algorithms, for data sharing that preserves autonomy of data sources and local access control policies.

## REFERENCES


[1] Chris Clifton, Murat KantarcioÇğlu, AnHai Doan, Gunther Schadow, Jaideep Vaidya, Ahmed Elmagarmid, and Dan Suciu. 2004. Privacy-preserving data integration and sharing. In *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM, 19–26.

[2] Ronald Fagin, Phokion G Kolaitis, Renée J Miller, and Lucian Popa. 2005. Data exchange: semantics and query answering. *Theoretical Computer Science* 336, 1 (2005), 89–124.

[3] Mehdi Haddad, Jovan Stevovic, Annamaria Chiasera, Yannis Velegrakis, and Mohand-Saïd Hacid. 2014. Access control for data integration in presence of data dependencies. In *International Conference on Database Systems for Advanced Applications*. Springer, 203–217.

[4] Alon Y Halevy. 2001. Answering queries using views: A survey. *The VLDB Journal* 10, 4 (2001), 270–294.

[5] Raghav Kaushik and Ravi Ramamurthy. 2011. Efficient Auditing for Complex SQL Queries. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data (SIGMOD '11)*. 697–708.

[6] Anastasios Kementsietsidis and Marcelo Arenas. 2004. Data sharing through query translation in autonomous sources. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. 468–479.

[7] Alexandros Labrinidis and Hosagrahar V Jagadish. 2012. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2032–2033.

# Entrepôts de données NoSQL orientés graphes

Hajer Akid
hajer.akid@etu.unistra.fr
Icube UMR7357, Pôle API
300 Bd Sébastien Brant, Université de Strasbourg
Illkirch 67400, France
REGIM-Lab LR11ES4, École Nationale d'Ingénieurs
de Sfax (ENIS), BP 1173, Université de Sfax
Sfax 3038, Tunisie

Gabriel Frey
g.frey@unistra.fr
Icube UMR7357, Pôle API
300 Bd Sébastien Brant, Université de Strasbourg
Illkirch 67400, France

Mounir Ben Ayed
mounir.benayed@ieee.org
REGIM-Lab LR11ES4, École Nationale d'Ingénieurs
de Sfax (ENIS), BP 1173, Université de Sfax
Sfax 3038, Tunisie
Département d'informatique et de communication, Faculté
des sciences de Sfax, Université de Sfax
Sfax 3038, Tunisie

Nicolas Lachiche
nicolas.lachiche@unistra.fr
Icube UMR7357, Pôle API
300 Bd Sébastien Brant, Université de Strasbourg
Illkirch 67400, France

De nos jours, les données collectées par les entreprises sont de plus en plus volumineuses, variées et générées avec une grande vitesse. Par ailleurs, les outils classiques de stockage, gestion et analyse de données ont atteint leurs limites. Ce phénomène connu sous le terme du Big Data [9], a poussé les entreprises à la recherche des alternatives permettant d'exploiter efficacement et rapidement les nouvelles sources de données. Dans ce contexte, les bases de données relationnelles, ont montré leurs limites principalement à cause du coût élevé de leurs passages à l'échelle et la rigidité de leurs schémas incompatibles avec les données non structurées [4, 13]. Pour répondre à ces problématiques, une nouvelle catégorie de bases de données NoSQL (Not Only SQL) est apparue dont les principaux avantages sont la lecture et l'écriture rapide de données, la flexibilité, la haute évolutivité, et le faible coût de stockage[12]. Le mouvement NoSQL regroupe quatre types de bases de données qui se distinguent par leurs modèles logiques : modèle orienté clé-valeur, orienté colonnes, orienté documents et orienté graphes. Durant ces dernières années, certains travaux se sont intéressés à l'utilisation des bases de données NoSQL pour mettre en place une nouvelle génération d'entrepôts de données massives et hétérogènes. Ces travaux ont revistié l'approche R-OLAP (Relationnel On-Line Analytical Processing) [10] basée sur le modèle relationnel . Certaines recherches ont défini de nouvelles approches basées sur le modèle orienté colonnes (C-OLAP)[1, 2, 6–8]. Dans [7], les chercheurs ont étudié l'impact de la normalisation des dimensions sur l'efficacité de l'entrepôt de données orienté colonnes. Les résultats montrent qu'à l'instar des bases de données relationnelles, le schéma en flocons de neige est beaucoup plus coûteux qu'un schéma en étoile. D'autres approches ont revisité le processus d'implantation des entrepôts de données à l'aide d'une approche basée sur le modèle orienté documents (D-OLAP)[5, 6]. Dans [5], l'effet de la normalisation des dimensions a été étudié. Les résultats montrent la dégradation de la performance des requêtes dans le cas d'un schéma en flocons de neige. Dans [3], une approche basée sur le modèle orienté graphes (G-OLAP) a été définie. L'objectif de ce travail est de redéfinir les opérateurs OLAP classiques en utilisant le langage Cypher du système de gestion de base de données orientée graphes Neo4j. Les travaux existant dans la littérature montrent la faisabilité de la mise en oeuvre d'un entrepôt de données NoSQL en utilisant les modèles orientés colonnes et orientés documents. Cependant, à notre connaissance, aucun travail n'a encore évalué la performance d'un entrepôt de données orienté graphes pour les deux variantes de schémas en étoile et en flocons de neige.

Dans ce travail, nous proposons une nouvelle approche G-OLAP qui permet de modéliser un entrepôt de données orienté graphes. Les deux modèles multidimensionnels en étoile et en flocons de neige ont été pris en compte. Dans notre approche, le méta-graphe d'un schéma multidimensionnel en étoile contient le nœud "fait" ayant comme label le nom du fait et regroupant les mesures sous forme de propriétés. Ces prorietés possèdent comme clé le nom de la mesure. Également, les dimensions sont représentées par un nœud de type "dimension" ayant le nom de la dimension comme label. Les paramètres et les attributs faibles sont des propriétés au sein du nœud de la dimension. Contrairement au modèle relationnel où les requêtes complexes nécessitent des jointures effectuées au moment du calcul, la liaison entre le fait et chaque dimension est traduite par un arc relatif à cette dimension. Par exemple, l'arc "Par_Date" relie le fait "Ventes" à la dimension "Date". L'arc "Par_Client" relie le fait "Ventes" à la dimension "Client". La puissance de ce modèle réside d'une part dans sa simplicité et d'autre part dans la facilité de navigation d'une entité à une autre. Les jointures coûteuses dans le cadre d'un modèle relationnel sont remplacées par des parcours simples du graphe minimisant ainsi le temps de réponse du aux

jointures. Le méta-modèle d'un schéma en flocons de neige est plus complexe que celui en étoile. Il contient plus de noeuds et d'arcs car les dimensions sont décomposées en plusieurs nœuds "paramètre" qui regroupent le paramètre et ses attributs faibles. Les paramètres d'une dimension sont reliés par des arcs et ordonnés suivant leurs niveaux de granularité formant ainsi une hiérarchie. Ainsi, le fait "Ventes" peut être observable "Par_Jour", "Par_Mois" et "Par_Année". Afin d'évaluer la performance des entrepôts de données proposés, nous avons commencé par la génération des données du benchmark TPC-DS [11]. Le schéma multidimensionnel du TPC-DS contient sept tables de faits et dix-sept tables de dimensions partagées ou non par les tables de faits. Parmi les données générées, nous avons util-isé le fait "Store_Sales" et ses dix dimensions. Parmi les dimensions liées au fait "Store_Sales", la dimension "Customer" a été principale-ment décomposée en plusieurs entités formant ainsi un schéma en flocons de neige. Nous avons transformé ce dernier en schéma en étoile en regroupant les différentes entités relatives aux clients au sein de la même dimension. Ensuite, nous avons appliqué notre ap-proche G-OLAP pour concevoir deux entrepôts de données orientés graphes ayant un schéma en étoile et en flocons de neige. Nous avons utilisé le système de gestion de base de données Neo4j pour les implémenter. En vue de comparer leurs performances avec les entrepôts de données relationnelles, nous avons utilisé l'approche R-OLAP pour créer deux entrepôts de données relationnels en étoile et en flocons de neige. Nous avons utilisé MariaDB comme base de données relationnelle. Après, nous avons utilisé les requêtes fournies par le benchmark TPC-DS qui interrogent la table de fait "Ventes" et ses dimensions pour comparer la performance de ces quatre entrepôts de données. Initialement, ces requêtes sont écrites en langage d'interrogation SQL. Nous les avons transformées en langage Cypher de la base Neo4j. Finalement, nous avons répété l'expérimentation pour d'autres volumes de données. Les résultats de nos expérimentations montrent la validité de notre approche. En plus, nous avons trouvé que pour certaines requêtes, les entrepôts de données orientés graphes sont beaucoup plus performants que les entrepôts de données relationnelles. Également, le schéma en flocons de neige n'est pas coûteux dans le cas d'un entrepôt de données orienté graphe.

Pour conclure, l'entrepôt de données orienté graphes est une solution prometteuse pour la modélisation, le stockage et l'analyse des données complexes. Dans nos futurs travaux de recherche, nous envisageons profitier de l'aptitude des bases de données orientées graphes à stocker et analyser des données complexes pour gérer des réseaux biologiques.

## REFERENCES


[1] Mohamed Boussahoua, Fadila Bentayeb, Omar Boussaid, and Nadia Kabachi. 2018. A Data Partitioning Optimization Approach for Distributed Data Warehouses on Column family NoSQL Systems. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), Springer, 54–60.

[2] Mohamed Boussahoua, Omar Boussaid, and Fadila Bentayeb. 2017. Logical schema for data warehouse on column-oriented NoSQL databases. In *International Conference on Database and Expert Systems Applications (DEXA)*. Springer, 247–256.

[3] Arnaud Castelltort and Anne Laurent. 2014. NoSQL Graph-based OLAP Analysis.. In *Knowledge Discovery and Information Retrieval (KDIR)*. 217–224.

[4] Rick Cattell. 2011. Scalable SQL and NoSQL data stores. *Acm Sigmod Record* 39, 4 (2011), 12–27.

[5] Max Chavalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2016. Document-oriented data warehouses: Models and extended cuboids, extended cuboids in oriented document. In *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 1–11.

[6] Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste, and Ronan Tournier. 2015. Implementing multidimensional data warehouses into NoSQL. International Conference on Enterprise Information Systems (ICEIS), 108–130.

[7] Khaled Dehdouh, Fadila Bentayeb, Omar Boussaid, and Nadia Kabachi. 2015. Using the column oriented NoSQL model for implementing big data warehouses. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 469–475.

[8] Khaled Dehdouh, Omar Boussaid, and Fadila Bentayeb. 2020. Big Data Warehouse: Building Columnar NoSQL OLAP Cubes. *International Journal of Decision Support System Technology (IJDSST)* 12, 1 (2020), 1–24.

[9] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management* 35, 2 (2015), 137–144.

[10] Daniel L Moody and Mark AR Kortink. 2000. From enterprise models to dimensional models: a methodology for data warehouse and data mart design.. In *Workshop on Design and Management of Data Warehouses (DMDW)*. ACM, 5–12.

[11] Raghunath Othayoth Nambiar and Meikel Poess. 2006. The making of TPC-DS. In *Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 1049–1058.

[12] Aaron Schram and Kenneth M Anderson. 2012. MySQL to NoSQL: data modeling challenges in supporting scalability. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*. ACM, 191–202.

[13] Asadulla Khan Zaki. 2014. NoSQL databases: new millennium database for big data, big users, cloud computing and its security challenges. *International Journal of Research in Engineering and Technology (IJRET)* 3, 15 (2014), 403–409.

# Micro-Environment Recognition in the Context of Mobile Sensing - A Holistic Approach

Hafsa El Hafyani
hafsa.el-hafyani@uvsq.fr
DAVID Lab, University of Versailles
Université Paris-Saclay

Karine Zeitouni
karine.zeitouni@uvsq.fr
DAVID Lab, University of Versailles
Université Paris-Saclay

Yehia Taher
yehia.taher@uvsq.fr
DAVID Lab, University of Versailles
Université Paris-Saclay

## ABSTRACT

The new mobile crowd sensing (MCS) paradigm leads to the generation of a large amount of data originated from different sources. The inhomogeneous nature of produced data turns the process of data analytics and knowledge extraction too much challenging and complicated. More specifically, mining such data needs special care and processing. In this paper, we focus on the activity recognition from GPS data. We advocate this process should benefit from the diverse data sources in the MCS context, including user's annotation and various ambient data collected by the sensors.

## KEYWORDS

Mobile Crowd Sensing, Activity Recognition, Data Mining, Spatio-temporal Series

## 1 INTRODUCTION

With the rapid advances of Internet of Things (IoT), more and more objects are now connected through GPS devices, ranging from humans, computers, animals and vehicles to the smallest devices. Therefore, there has been a large amount of generated trajectory data that needs to be stored, processed, and analyzed. Recently, research initiatives related to trajectory data focus mainly on data mining and knowledge extraction. However, the application of traditional methods needs clean, pruned and ready-to-use trajectories. Yet, the real-world trajectory data are usually imprecise due to noise. Therefore, a framework that transforms raw trajectory data to enriched and noise-resistant data is more needed than ever.

One of the applications of IoT is the new paradigm called Mobile Crowd Sensing (MCS), which empowers volunteers to contribute data acquired by their personal sensor-enhanced mobile devices. Polluscope [2], a french project funded by ANR and deployed in Île-de-France, is a typical use case study based on MCS. It aims at getting insight constantly on individual exposure to pollution everywhere (indoor and outdoor), while enriching the traditional monitoring system with the collected data by the crowd. Data derived from MCS have, however, a high uneven density in time and space, which makes mining such inhomogeneous data far from being direct [7].

In this paper, we will tackle the problem of mining trajectory data collected in the context of MCS. In particular, we aim at deriving the user's activities from its trajectory as well as other sources, among which Points of Interest (PoIs), but also the ambient air data series. Indeed, activity recognition using mobile sensors empowered with GPS chipsets is a trending topic [8], but it rarely exploits other contextual data.

We will start by defining the structure of data collected to fulfill those objectives. The problematics of activity recognition are then presented in section 3. The challenges encountered while tackling those problematics will be the subject of section 4.



## 2 CONTEXT

The MCS paradigm allows the use of opportunistic air quality monitoring, where emerging low-cost and lightweight air pollution sensors are fetched on pedestrians, cyclists, or vehicles to perform air quality monitoring. This model enables insights at the finest level along the participant daily trajectories, thereby allowing to capture local variability and peaks of pollution. This leads to a massive trajectories along with sensor data series. Due to the complexity of their spatial and the temporal dimensions (interdependence, skewness in the spatial and temporal distributions, noise, etc.), analyzing spatio-temporal data is not simple. In fact, direct use of the state-of-the-art methods, such as data mining, is far from being straightforward.

In the context of our work, we aim at developing a framework based on data mining approaches, techniques, and technologies that use geo-dated series as input to extract knowledge from participants trajectories and learn meaningful patterns. In this work, we focus on activity recognition in order to annotate the trajectories with, e.g., transportation mean used for mobility, indoor or outdoor micro-environment of the participant (Home, Work, Streets, Parks, etc.) [6]. We also look for the possible correlations between ambient air quality profile and the micro-environment.

## 3 METHODOLOGY

Figure 1 provides an extensive representation of the workflow of our framework. The model focuses mainly on the detection of stops and moves within trajectory data, and on their enrichment.

### 3.1 Segmentation

The trajectory segmentation procedure divides each trajectory into sub-trajectories called segments. Trajectories can be viewed as sequence of stops and moves. A stop sequence implies that the moving object is static at some location, while a move sequence connects two consecutive updates by a mobility mean. The detection of stops in a moving object trajectory poses a critical problem in the activity recognition process. Data collected by mobile sensors are not always precise due to the noise caused mainly by the limitations

, ,

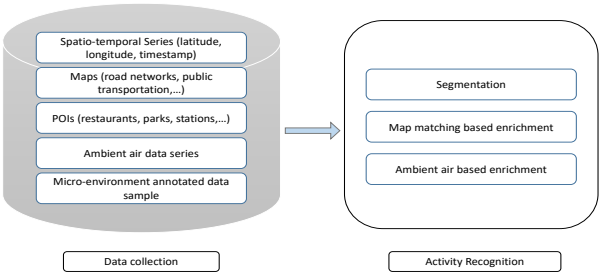


**Figure 1: Workflow of the activity recognition process**

of positioning systems. For this reason, we use DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [4], one of the most robust clustering algorithm, to group points sequences with high density into one cluster. The created clusters absorb all points, including noise, in its neighborhood.

With the application of DBSCAN, two points that do not belong to the same time range, may fit into the same cluster, since DBSCAN is a not time-aware process. We conduct a post processing technique to improve our framework performance. The post-processing technique is based on defining a threshold for the time lag between two updates, and another one for the number of points in the cluster. If the duration is higher than the threshold and the minimal density is satisfied, we split the cluster into N clusters.

### 3.2 Map matching based enrichment

Map matching is the procedure of adding contextual data to the raw trajectory data [5]. The GPS data are matched with road networks, public transportation system, and POIs within the city. This process will enhance the detection of user's micro-environment (e.g., restaurant, home, train station) when we match GPS data with different related maps. And thus, we establish a dimension multiplicity process depicted by the underlying micro-environment (e.g., indoor and outdoor spaces), contextual data (e.g., POIs), as well as sensors data. Operations on such data using spatial OLAP will allow us to get insights of information at different levels (i.e., relation between transportation mean and hygrometry) [3].

### 3.3 Ambient air based enrichment

The stop and move detection model contributes to the detection of the user's micro-environment as well. Our framework tries to perform supervised and unsupervised approaches to classify multivariate time series (e.g. pollutant measurements) by micro-environment location. A possible correlation between ambient air data series and user's whereabouts is to be examined. One of the strengths of our work is the use of this correlation, if existed, in order to enhance the qualification of micro-environment detection.

### 3.4 Micro-environment annotated data

A sample of collected data is transportation mean annotated. We use that sample as a ground truth in the process of supervised learning. In addition, data contain pollution-related events (e.g., cooking, opening window) which care to check the effect of human behaviors on ambient air in an indoor location.

## 4 OUTLOOK

Research on activity recognition and detection of micro-environment with the contribution of ambient air data series could continue on several directions. In this section, we present the outline that we predict to consider in the context of this work.

- **Semantic Enrichment and Algorithms Optimization:** Activity recognition already allows to transform raw trajectory data to semantic trajectories, i.e., an annotated sequence micro-environments. We need to go further by deriving the semantic of the moves (i.e. walking, car, bus, etc.) where again the signature of the environmental data series can help. Other annotations might be of interest, typically some events where the air quality is a marker (i.e., opening a window, passing by a bus exhaust, lighting a cigarette, etc.). To do so, we need to collect for each event or environment, its characteristics in term of ambient air quality. Furthermore, We aim at developing a near-real time algorithm that classifies the input data stream into a stop or a move segment as it arrives to the system, and extracts its micro-environment.
- **Data Analytics and Visualization:** Data visualization enables the detection of ambient air data series signature on micro-environment. We take advantage from an online visualization open source scheme [1] to visualize constantly users activities, ambient air data series, change of environment, and abnormalities. Coupling data analysis spatial OLAP queries will allow to interactively explore the data on different dimensions and scales. We also intend to apply data mining techniques to assist in suggest the grouping in the data cube. A promising approach is biclustering which interest is to simultaneously cluster rows and columns of a data cube.

## 5 CONCLUSION

In this paper, we propose a strategy to convert raw trajectory data into knowledge with the support of map matching based enrichment and ambient air based enrichment. More specifically, we present a holistic approach on detecting and extracting micro-environment space from GPS trajectory data by processing ambient air data collected by mobile crowd sensing technologies.

## 6 ACKNOWLEDGMENT

This work is supported by the grant ANR-15-CE22-0018 Polluscope of the French National Research Agency (ANR).

## REFERENCES

[1] [n. d.]. Grafana Labs. https://grafana.com/. Last accessed 2 September 2019.
[2] [n. d.]. POLLUSCOPE project. http://polluscope.uvsq.fr. Last accessed 2 September 2019.
[3] E. Bernier, P. Gosselin, T. Badard, and Y. BÃľdard. 2009. Easier surveillance of climate-related health vulnerabilities through a Web-based spatial OLAP application. *International Journal of Health Geographics* 8 (2009).
[4] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96)* (1996), 226–231.

[5] C. Parent and S. Spaccapietra et al. 2013. Semantic Trajectories Modeling and Analysis. *ACM Computing Surveys (CSUR)* (2013).
[6] K. Sila-Nowicka and P. Thakuriah. 2019. Multi-sensor movement analysis for transport safety and health applications. *PloS one* 14 (2019), e0210090.
[7] H. Su, K. Zheng, K. Zeng, J. Huang, and X. Zhou. 2014. STMaker: a system to make sense of trajectory data. *Proceedings of the VLDB Endowment* 7 (2014), 1701–1704.
[8] Y. Zheng, L. Liu, L. Wang, and X. Xie. 2008. Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. *International World Wide Web Conference Committee (IW3C2)* (2008).

# Analysing Conservation Data at the BnF

Alaa Zreik
University of Versailles Saint-Quentin-En-Yvelines
Versailles 78000
alaa.zreik@ens.uvsq.fr

Zoubida Kedad
University of Versailles Saint-Quentin-En-Yvelines
Versailles 78000
zoubida.kedad@uvsq.fr

## ABSTRACT

The conservation-restoration process at the French National Library (BnF) aims to conserve the physical state of the BnF's documents and prevents it from becoming out-of-order. The goal of our work is to provide support in order to select the documents which have to undergo a conservation / restoration process and to identify the ones to which priority should be given. To this end, we propose to analyze the history of conservation / restoration data for the documents. In this paper, we present an exploratory data analysis at the BnF in collaboration with BnF's experts on the population of out-of-order documents, aiming to discover the causes of their bad states. On the other hand, we analyzed the documents treatments data in order to find specific types of trajectories that characterize a high number of out-of-order documents. Finally, we present a clustering method based on HDBScan using a custom trajectory similarity function to detect the characteristics of out-of-use documents.

## KEYWORDS

Conservation-restoration, Data analysis, Predictive analysis

## 1 INTRODUCTION

A document at the BnF is exposed to external actions that could change its state; the state of a document can be one of the three followings: communicable, non communicable and out-of-use. Communicable documents are the documents that are in good physical condition and can be requested by readers. Non communicable documents are the documents requiring certain treatments before becoming communicable. The out-of-use group contains documents that are in poor physical state and can not be requested by readers. Two types of actions exist, degradations and treatments. The degradations are the actions that deteriorate the physical state of the documents and change it to non-communicable or out-of-order state. A degradation can be natural or unnatural, factors such as humidity and tearing, respectively. On the other hand, the treatments are the actions that can improve the state of the documents such as bookbinding. The libraries consider the out-of-use documents as being lost, due to the fact that the importance of a document lies in the possibility of transferring it to the public. According to that, we started a lengthy exploratory data analysis (EDA) at the BnF, aimed to understand the basic characteristics of the documents, treatments, degradations and communications. After the EDA process, our goal is to study the documents treatments cycles as trajectories. Generally speaking, a trajectory is a set of nodes (places, states) at specific moments, with links between them (streets, actions). In our context, the document's treatments are considered as a trajectory where the nodes represent the different treatments. The link between two treatments t1 and t2 occurring at the dates d1 and d2 respectively are the number of the document's requests by readers and the time between d1 and d2. Finally, in order to find the most frequent characteristics in the trajectories of out-of-use documents, we will use clustering algorithms on the documents trajectories. In our first experiment, we have used an algorithm relying on hierarchical density-based spatial clustering with noise (HDBScan) with a custom predefined trajectory similarity function.

The paper is structured as follows: in section 2, we present some results of the EDA process. In section 3 we define our custom similarity function and introduce the trajectories clustering method. Finally section 4 introduces the future works for predicting conservation actions by analyzing the documents trajectories.



## 2 EXPLORATORY DATA ANALYSIS

In this section, we present some results of the analysis on the characteristics of cultural objects at the BnF. the goal of this analysis is to determine the percentage of each type of objects at the BnF, the frequency of requests for documents by readers and the percentage of out-of-use documents per year.

### 2.1 Distribution by type

We have determined the distribution of objects according to their type and the percentages of each type, which can help us to identify a specific type on which to deepen our study. We obtained the following result:

- `Printed matter`: 51%.
- `Digital objects`: 17%.
- `Sound records`: 8%.
- `Others`: 24%

After this first result, we focused on a specific category of documents, namely printed documents, which represents the largest category at the BnF.

### 2.2 Readers requests

Another dimension of analysis is the number of requests of documents.Information about the readers requests of documents since 2016 is available. By analyzing this information, we obtained some results which shows the interest of readers in the XXI century and a slight increase in interest between 1841 and 1860.

### 2.3 Out of order documents

In terms of communicability, a document can be in one of the three following states: communicable, non-communicable and out of order; these states correspond to the physical state of the document.

The out of order group contains all documents in very poor condition due to their age or to a heavy use. Out of order documents can be considered as a loss for libraries because the importance of a document lies in the possibility of transferring it to the public. At the BnF, there are currently about 170,000 out-of-order documents dating from several periods, from the XV century to the present day. One of our goals is to detect the characteristics of these documents in order to better understand the causes of degradation.

By analyzing the BnF databases, we found out that the percentage of out-of-order documents is high for the ones dating from the early fifteenth century and the nineteenth century.

## 3 TRAJECTORIES CLUSTERING

After considering the treatments cycles of documents as trajectories, we have proposed the use of a clustering method that rely on the hierarchical density-based spatial clustering of applications with noise (HDBScan). The idea is to cluster similar trajectories (i.e. documents treatments cycles) and to analyze the clusters having a high percentages of out-of-order documents. Clusters analysis can provide us with relevant information on the causes of degradation, such as the most frequent treatments in the life cycles of out-of-order documents.

### 3.1 Trajectories

There are nearly 100 different treatments at the BnF. Each treatment consists of one or more processes. There are nearly 400 different processes at the BnF. In this paper, a treatment is denoted as $T'x$, and a process is denoted as $P'x$, where $x$ is the unique identifier of the treatment and the process. For example, consider a document which has undergone two treatments $T_1$ and $T_{15}$ and each treatment is composed of $[P_3, P_{400}, P_{510}]$ and $[P_3, P_{20}, P_{300}]$ respectively. We define the trajectory as the whole union, and the trajectory of the document in the example will be the sequence of processes $Tr = [P_3, P_3, P_{20}, P_{300}, P_{400}, P_{510}]$.

### 3.2 Similarity

Every clustering method requires a predefined distance function. Several distance functions have been proposed for trajectories clus-tering as described in [1]. we define the distance between two trajectories $T_1$ and $T_2$ as follows:

$$Distance(Tr_1, Tr_2) = 1 - SIM(Tr_1, Tr_2) \quad (1)$$

Where SIM is the function calculating the similarity between two trajectories. The value of *SIM* is equal to 1 if the two trajectories are the same, and 0 if they are completely different.

$$SIM(Tr_1, Tr_2) = \frac{|Tr_1 \cap Tr_2|}{|Tr_1 \cup Tr_2|} \quad (2)$$

### 3.3 Clustering method

We have applied our clustering approach, which relies on HDBScan, with several values of the *mindistance* parameter, starting from 0.8. After each clustering execution, we decreased the parameter by 0.05 and we performed the next execution on a selected subset of clusters. For this reason, we have defined three types of clusters, namely Selected, Dropped and toExplore clusters.

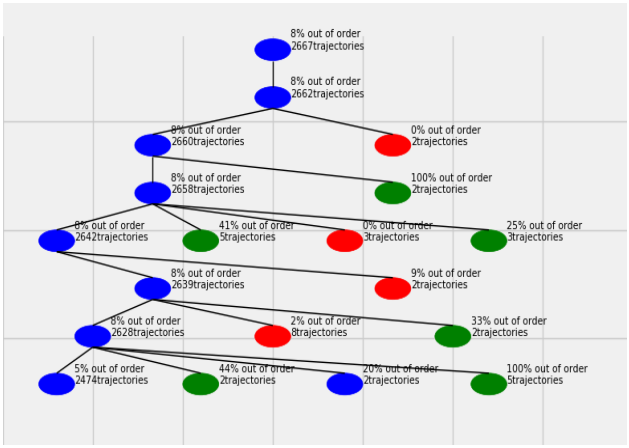


**Figure 1: DBScan Hierarchical clustering results**

*3.3.1 Selected clusters.* A selected cluster is a cluster where the percentage of out-of-use documents is more than 80%, or a cluster having a number of trajectories less than 10 with a percentage of out of use documents greater than 50%. This type is represented by the green color in the figure 1.

*3.3.2 Dropped clusters.* A dropped cluster is a cluster where the percentage of out of use documents is less than 20%. This type is represented by the red color in the figure 1.

*3.3.3 ToExplore clusters.* A toExplore cluster is a cluster that should move to the next level of clustering. Clusters in this group have more than 10 trajectories and a percentage of out of use between 20% and 80%. This type is represented by the blue color in the figure 1.

## 4 FUTURE WORK

The next step will be to analyze the selected clusters and create patterns that will contain the most frequent treatments. These patterns will be used to detect vulnerable documents, which will be the ones similar to some of the pre-created patterns. We will then adapt the clustering method to take into account other dimensions and characteristics of documents, such as information about document communications to readers.

## REFERENCES

[1] Zhang Zhang, Kaiqi Huang, Tieniu Tan, et al. 2006. Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes.. In *ICPR (3)*. Citeseer, 1135–1138.

# 7 Résumés des démonstrations

# ARIANE : la Gouvernance des Données comme Accélérateur de Conformité au Règlement Général sur la Protection des Données

Mehdi BENTOUNSI
Umanis, Levallois-Perret, France
mebentounsi@umanis.com

Edouad CANTE
Blueway Software, Lyon, France
ecante@blueway.fr

Daniel COYA
Blueway Software, Lyon, France
dcoya@blueway.fr

Patrice DARMON
Umanis, Levallois-Perret, France
pdarmon@umanis.com

Arnaud DE CHAMBOURCY
Umanis, Levallois-Perret, France
adechambourcy@umanis.com

Gisèle GNOKAM
Umanis, Levallois-Perret, France
ggnokam@umanis.com

**RÉSUMÉ**

Assurer la conformité au règlement général sur la protection des données (RGPD) passe par la mise en place de la protection de la vie privée dès la conception des processus métiers des organisations (privacy by design). Il est par conséquent nécessaire de prendre en compte les contraintes liées à l'usage des données à caractère personnel dans le plan d'urbanisme des systèmes d'informations (SI). La démonstration présente ARIANE, une plateforme intégrée de gouvernance des données à caractère personnel. ARIANE permet d'industrialiser la protection de la vie privée en constituant un référentiel unique de personnes physiques.

*Conference'17, July 2017, Washington, DC, USA* 2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00 https://doi.org/10.1145/nnnnnnn.nnnnnnn

# Spade: A Modular Framework for Analytical Exploration of RDF Graphs

Yanlei Diao[1,2] Paweł Guzewicz[1,3] Ioana Manolescu[1,3] Mirjana Mazuran[1,3]
[1] LIX (UMR 7161, CNRS and Ecole polytechnique), France
[2]University of Massachusetts Amherst
[3] Inria, France

yanlei.diao@polytechnique.edu, {pawel.guzewicz,ioana.manolescu,mirjana.mazuran}@inria.fr

## ABSTRACT

RDF data is complex; exploring it is hard, and can be done through many different metaphors. We have developed and propose to demonstrate Spade, a tool helping users discover meaningful content of an RDF graph by showing them the results of *aggregation (OLAP-style)* queries automatically identified from the data. Spade chooses aggregates that are *visually interesting*, a property formally based on statistic properties of the aggregation query results.

While well understood for relational data, such explo-ration raises multiple challenges for RDF: facts, dimensions and measures have to be *identified* (as opposed to known be-forehand); as there are more candidate aggregates, assessing their interestingness can be very costly; finally, *ontologies* bring novel specific challenges but also novel opportunities, enabling *ontology-driven exploration* from an aggregate ini-tially proposed by the system.

Spade is a *generic, extensible framework*, which we instan-tiated 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 inter-estingness functions; (*iii*) ontology-based interactive explo-ration, and (*iv*) efficient early-stop techniques for estimating the interestingness of an aggregate query.

The demonstration will comprise interactive scenarios on a variety of large, interesting RDF graphs.

**PVLDB Reference Format:**
Y. Diao, P. Guzewicz, I. Manolescu, M. Mazuran. Spade: A Mod-ular Framework for Multi-Dimensional RDF Exploration. *PVLDB*, 12(12): xxxx-yyyy, 2019.
DOI: https://doi.org/10.14778/xxxxxxx.xxxxxxx


*Proceedings of the VLDB Endowment,* Vol. 12, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/xxxxxxx.xxxxxxx

# Neo4Tourism - A framework for Graph Data Analysis on Tourism

Gaël Chareyron
Ugo Quelhas
Nicolas Travers
firstname.lastname@devinci.fr
Leonard de Vinci, Research Center
Paris La Défense, France

## ABSTRACT

Digital Tourism has become a popular mean for analyzing tourists' behavior over time. It essentially enhances traditional ground stud-ies with massive data analysis to validate models. In such environ-ment, tourism actors are faced with the need to deeply understand tourists' circulation both quantitively and qualitatively. Thus, deal-ing with huge volumes of data from social networks dedicated to tourism experience is a way to cope with this issue.

We propose in this paper the `Neo4Tourism` framework based on a graph data model specialized in digital tourism analysis. Our model is dedicated to tourists' circulation and aims at simulating tourists' behavior. In this demonstration we discuss how our system integrates data from TripAdvisor in a *Neo4j* graph database, also how it produces circulation graphs, enhance graphs manipulations and deep tourists' analysis by identifying centralities of locations.

## KEYWORDS

Graph databases, Digital Tourism, Neo4j, Geodesy

## ACKNOWLEDGMENTS

This work has been funded by the research agreement "Big Data & Tourism" in collaboration with Bordeaux Métropole, Tourism Office of Bordeaux and the *Caisse des Dépôts et Consignations* (CDC).

# DISPERS: Securing Highly Distributed Queries on Personal Data Management Systems

Julien Loudet[1,2,3]
[1]Cozy Cloud, France
contact@cozycloud.cc

Iulian Sandu-Popa[3,2]
[2]INRIA Saclay, France
<fname.lname>@inria.fr

Luc Bouganim[2,3]
[3]University of Versailles, France
<fname.lname>@uvsq.fr

Grâce à de nouvelles initiatives [4] ainsi qu'à de nouvelles réglementations [6], nous pouvons accéder aux données personnelles que les entreprises et agences gouvernementales ont collectées sur nous. En parallèle, de plus en plus de *Systèmes de Gestion de Données Personnelles (SGDP)* voient le jour, à la fois dans l'industrie [5] et dans le milieu académique [1]. Leur objectif est d'offrir une plateforme de gestion de données qui permet aux utilisateurs de facilement conserver, au même endroit, n'importe quelle information : (i) directement générée par un de ses appareils ou (ii) issue d'une interaction avec un service (par exemple, des données de santé ou bancaires). Les utilisateurs peuvent ensuite utiliser leur SGDP dans leur propre intérêt ou dans celui d'une communauté. Ainsi, le paradigme du SGDP promet de créer de nouvelles applications centrés autour de la donnée personnelle. Un exemple proéminent de ces nouveaux usages a trait aux calculs distribués qui impliquent un grand nombre de SGDP (par exemple, des études participatives).

Cependant, ces perspectives excitantes ne doivent pas éclipser les enjeux de sécurité soulevés par ce nouveau paradigme : conserver la totalité de sa vie numérique au même endroit augmente proportionnellement l'impact d'une fuite. Il devient donc risqué de centraliser l'ensemble des données des utilisateurs sur des serveurs puissants puisque ces derniers deviennent des cibles de choix pour des attaquants [7]. De surcroît, une solution centralisée serait antinomique dans ce contexte puisque les données sont naturellement distribuées chez les utilisateurs.

Heureusement, les « environnements d'exécution de confiance » (*"Trusted Execution Environment" — TEE*) [2, 9] deviennent de plus en plus présents et en les combinant avec les SGDP nous obtenons une plateforme de calcul robuste. Néanmoins, comme aucune mesure de sécurité ne peut être considérée comme inviolable, nous ne pouvons exclure l'hypothèse qu'une partie des SGDP du système soit corrompue, que ces SGDP collaborent, voire même qu'ils soient indétectables des autres SGDP honnêtes [3]. Nous supposons dans ces travaux qu'un SGDP est donc sécurisé par un TEE, qu'il peut être corrompu sans que cela soit détectable, qu'il offre une excellente connectivité et disponibilité, et qu'il puisse établir des connexions en pair-à-pair avec d'autres SGDP. Ainsi, nous envisageons une architecture entièrement distribuée de SGDP dans laquelle les participants peuvent créer de larges communautés, contribuer avec leurs données personnelles et interroger l'ensemble de ces données. Dans ce contexte, une question importante doit être traitée : comment interroger cette masse de données distribuée de manière pertinente, efficace et tout en respectant la vie privée des participants ?

DISPERS respecte les trois principes suivants pour interroger les données des utilisateurs de manière sécurisée : la *dispersion des connaissances* pour protéger les données « au repos », la *compartementalisation des tâches* pour protéger les données « en utilisation », et l'« *aléa imposé* », basé sur nos précédents travaux SEP2P [8], pour assigner les tâches aux SGDP de manière aléatoire et vérifiable. Ces travaux se concentrent exclusivement sur ce dernier principe, proposant une solution générique, efficace et capable de passer à l'échelle même en étant confronté à un important nombre de SGDP corrompus qui collaborent.

À notre connaissance, DISPERS est le premier protocole qui adresse cette problématique. Il va plus loin que des solutions classiques (voir [8]) qui ont été proposées dans des domaines tels que les Tables de Hachage Distribuées Securisées [12], les Calculs Sécurisés Mulitpartites [10], ou l'agrégation de données distribuées se basant sur du matériel sécurisé [11]. Il est également à noter que les protocoles de consensus et les systèmes résistants aux fautes Byzantines répondent à des problématiques différentes et ces solutions améliorent généralement la disponibilité et l'intégrité des données au détriment de la confidentialité [13].

## RÉFÉRENCES

[1] Serge Abiteboul, Benjamin André, and Daniel Kaplan. 2015. Managing your digital life. *Comm. of the ACM* 58, 5 (2015).
[2] Nicolas Anciaux, Philippe Bonnet, Luc Bouganim, Benjamin Nguyen, Philippe Pucheral, Iulian Sandu Popa, and Guillaume Scerri. 2019. Personal Data Management Systems : The security and functionality standpoint. *Information Systems* 80 (2019).
[3] Yonatan Aumann and Yehuda Lindell. 2007. Security against covert adversaries : Efficient protocols for realistic adversaries. In *Theory of Cryptography*.
[4] Blue Button. 2017. Find Your Health Data https://www.healthit.gov/topic/health-it-initiatives/blue-button.
[5] Cozy Cloud. 2018. Your digital home. https://cozy.io/en.
[6] European Parliament. 2016. General Data Protection Regulation. Law.
[7] Troy Hunt. 2018. ';–Have I been pwned? Largest and recent breaches. https://haveibeenpwned.com/.
[8] Julien Loudet, Iulian Sandu-Popa, and Luc Bouganim. 2019. SEP2P : Secure and Efficient P2P Personal Data Processing. In *EDBT*.
[9] Christian Priebe, Kapil Vaswani, and Manuel Costa. 2018. EnclaveDB : A Secure Database Using SGX. In *IEEE Symposium on Security and Privacy*.
[10] Eyad Saleh, Ahmad Alsa'deh, Ahmad Kayed, and Christoph Meinel. 2016. Processing over encrypted data : between theory and practice. *ACM SIGMOD Record* 45, 3 (2016), 5–16.
[11] Quoc-Cuong To, Benjamin Nguyen, and Philippe Pucheral. 2016. Private and scalable execution of SQL aggregates on a secure decentralized architecture. *ACM Transactions on Database Systems (TODS)* 41, 3 (2016), 16.
[12] Qiyan Wang and Nikita Borisov. 2012. Octopus : A secure and anonymous DHT lookup. In *Distributed Computing Systems (ICDCS)*.
[13] Jian Yin, Jean-Philippe Martin, Arun Venkataramani, Lorenzo Alvisi, and Mike Dahlin. 2003. Separating Agreement from Execution for Byzantine Fault Tolerant Services. In *ACM Symposium on Operating Systems Principles (SOSP)*. 15. https://doi.org/10.1145/945445.945470

# Algorithmes distribués pour la recherche de séries temporelles similaires

Oleksandra Levchenko
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
oleksandra.levchenko@inria.fr

Boyan Kolev
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
boyan.kolev@inria.fr

Djamel-Edine Yagoubi
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
djamel-edine.yagoubi@inria.fr

Dennis Shasha
Dep. of Computer Sc., New York University
New-York, USA
shasha@cs.nyu.edu

Themis Palpanas
University of Paris
Paris, France
themis@mi.parisdescartes.fr

Reza Akbarinia
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
reza.akbarinia@inria.fr

Florent Masseglia
LIRMM, Univ Montpellier, Inria, Montpellier, France
Montpellier, France
florent.masseglia@inria.fr

## ABSTRACT

Avec les progrès des appareils de mesure, le besoin d'algorithmes efficaces et évolutifs augmente pour fusionner les séries temporelles qui en résultent. Les capteurs produisent des milliers, voire des milliards de séries temporelles, de sorte que la première étape de la fusion consiste souvent à trouver des séries temporelles similaires. Les applications comprennent les stratégies d'arbitrage statistique dans le domaine financier et la détection des tremblements de terre dans les données sismiques.

Pour traiter un si grand nombre de séries chronologiques, les algorithmes demandent une indexation très performante. La création d'un index sur des milliards de séries temporelles en utilisant des approches centralisées traditionnelles prend beaucoup de temps.

Une possibilité intéressante pour améliorer les performances de la construction d'index et de la recherche de similitude sur des ensembles aussi massifs de séries chronologiques consiste donc à tirer parti de la puissance de calcul des systèmes distribués et des architectures parallèles. Toutefois, une parallélisation naïve des techniques existantes sous-exploiterait la puissance de calcul disponible. Nous avons mis en œuvre des algorithmes parallèles pour deux approches de pointe permettant de construire des index et de fournir une recherche de similitude sur de grands ensembles de séries temporelles en répartissant soigneusement la charge de travail. Notre solution tire parti de la puissance de calcul des systèmes distribués en utilisant des architectures parallèles, en l'occurrence Spark.

Cette démonstration utilise des données financières et sismiques pour montrer comment deux algorithmes de pointe construisent des index et répondent à des requêtes de similitude en utilisant Spark. Le public de la démonstration pourra choisir des séries temporelles requêtes, voir comment chaque algorithme se rapproche de ses voisins les plus proches, et comparer les temps de réponse dans un environnement parallèle.

## KEYWORDS

Dirichlet Process Mixture Model, Clustering, Parallelism

## 1 ACKNOWLEDGEMENTS

The research leading to these results has received funds from the European Union's Horizon 2020 Framework Programme for Research and Innovation, under grant agreement No. 732051.

# ExplIQuE : Exploration Interactive de Bases de Données en SQL

Marie Le Guilly
marie.le-guilly@liris.cnrs.fr
Univ Lyon, INSA Lyon, LIRIS (UMR 5205 CNRS)
Villeurbanne, France

Jean-Marc Petit
jean-marc.petit@liris.cnrs.fr
Univ Lyon, INSA Lyon, LIRIS (UMR 5205 CNRS)
Villeurbanne, France

Marian Scuturici
marian.scuturici@liris.cnrs.fr
Univ Lyon, INSA Lyon, LIRIS (UMR 5205 CNRS)
Villeurbanne, France

Ihab Ilyas
ilyas@uwarteloo.ca
University of Waterloo
Canada

## 1 CONTEXTE

ExplIQuE (*Exploration Interface with Query extensions*) est une application ayant pour but d'aider les utilisateurs de bases de données, qui viennent juste d'apprendre le SQL et/ou qui ne sont pas familiers avec leurs données. En effet, le volume de données stockés est en constante augmentation, créant de nouveaux besoins chez les utilisateurs, pour stocker, interroger, et analyser des données. Dans ce contexte, et en fonction des connaissances de l'utilisateur sur ses données, la requête à poser sur la base n'est pas forcément claire, et peut donc être difficile à traduire directement en SQL : le but d'ExplIQuE est alors d'aider l'utilisateur à formuler de telles requêtes que nous qualifions d'*imprécises*. De telles requêtes sont courantes dans des contextes exploratoires, où il est nécessaire de bien connaître les données, et d'essayer plusieurs requêtes différentes avant d'atteindre les données désirées. Ainsi, dans un tel cas, un utilisateur va commencer par quelques requêtes pour explorer et mieux comprendre les données. Il peut ensuite commencer par une requête assez générale, puis la raffiner de manière itérative, jusqu'à parvenir à répondre à la question initialement posée.

Ce processus itératif peut ne pas être aisé pour l'utilisateur, qui peut ne pas savoir par où commencer, ou comment affiner sa requête pour qu'elle ne renvoie que les tuples qui l'intéressent. Plusieurs solutions ont été proposées pour faciliter cette étape du processus d'exploration [3]. ExplIQuE a pour but d'assister l'utilisateur dans la phase de raffinement, en proposant des extensions correspondant à des complétions de la requête courante, sous la forme de prédicats de sélection supplémentaires. Ces extensions, générées en fonction des données de la base, permettent alors de compléter automatiquement la requête, et d'obtenir une nouvelle requête pour laquelle des extensions peuvent également être calculées. L'interface propose de l'aide pour choisir l'extension la plus appropriée parmi celles proposées, en se basant sur différentes métriques et visualisations.

Pour résumer, les trois principaux objectifs d'ExplIQuE sont les suivants :

— Aider les utilisateurs à mieux comprendre leurs données, en s'intéressant à des attributs spécifiques, en observant comment les résultats de leur requête peuvent être divisés, et en les visualisant.



— Débloquer les utilisateurs qui ne savent pas par où commencer, en leur proposant des suggestions pour spécifier une requête très générale.

— Proposer une aide sémantique, là où la plupart des éditeurs SQL ne proposent qu'une aide syntaxique, qui ne se base pas sur le contenu de la base de données.

ExplIQuE est conçu comme une interface pouvant se connecter à n'importe quelle base de données, et proposant les fonctionnalités relatives à l'extension de requêtes en plus des fonctionnalités classiques des éditeurs SQL. L'objectif de la démonstration est de montrer ces fonctionnalités sur une base de données scientifiques, et en posant des questions imprécises à l'audience de la démonstration, qui peut alors chercher à y répondre en utilisant ExpliQuE.

## 2 PRÉSENTATION DU SYSTÈME

La principale fonctionnalité d'ExpliQuE est la proposition d'extension à une requête SQL donnée (voir [1] pour plus de détails). Ces extensions sont des clauses de sélection supplémentaires, qui s'insèrent directement dans la clause *Where* de la requête considérée. Pour une même requête, plusieurs extensions sont proposées, qui couvrent les résultats de la requête initiale. Ainsi, ces extensions permettent à l'utilisateur de mieux comprendre comment peuvent se répartir les tuples de sa requête initiale, et donc de mieux comprendre sa base de données. Dans cette optique, les extensions sont calculées en deux étapes :

— Dans un premier temps, les tuples de la requête à étendre sont regroupés par un algorithme de clustering, afin de regrouper les tuples les plus similaires. En effet, les tuples recherchées par l'utilisateur ont généralement des caractéristiques communes, que le clustering peut permettre d'identifier. L'algorithme utilisé est *k-means* [2] : l'interface permet de tester différentes valeurs de $k$, pouvant être paramétrées par l'utilisateur, comme étant le nombre d'extensions à obtenir. Cela permet à l'utilisateur d'avoir un contrôle sur le nombre de clusters et donc d'extension, dans le cas où il a des connaissances du domaine pouvant influer sur la valeur à choisir.

— Chaque tuple est ensuite labellisé par le numéro du cluster auquel il appartient. Un arbre de décision est alors utilisé, pour discriminer les différents clusters. La taille de l'arbre est limitée, pour obtenir autant de feuilles que de clusters, afin de garder une cohérence entre le nombre d'extensions demandé et le nombre de résultats obtenus. En effet, chaque feuille de l'arbre donne naissance à une extension, en parcourant l'arbre

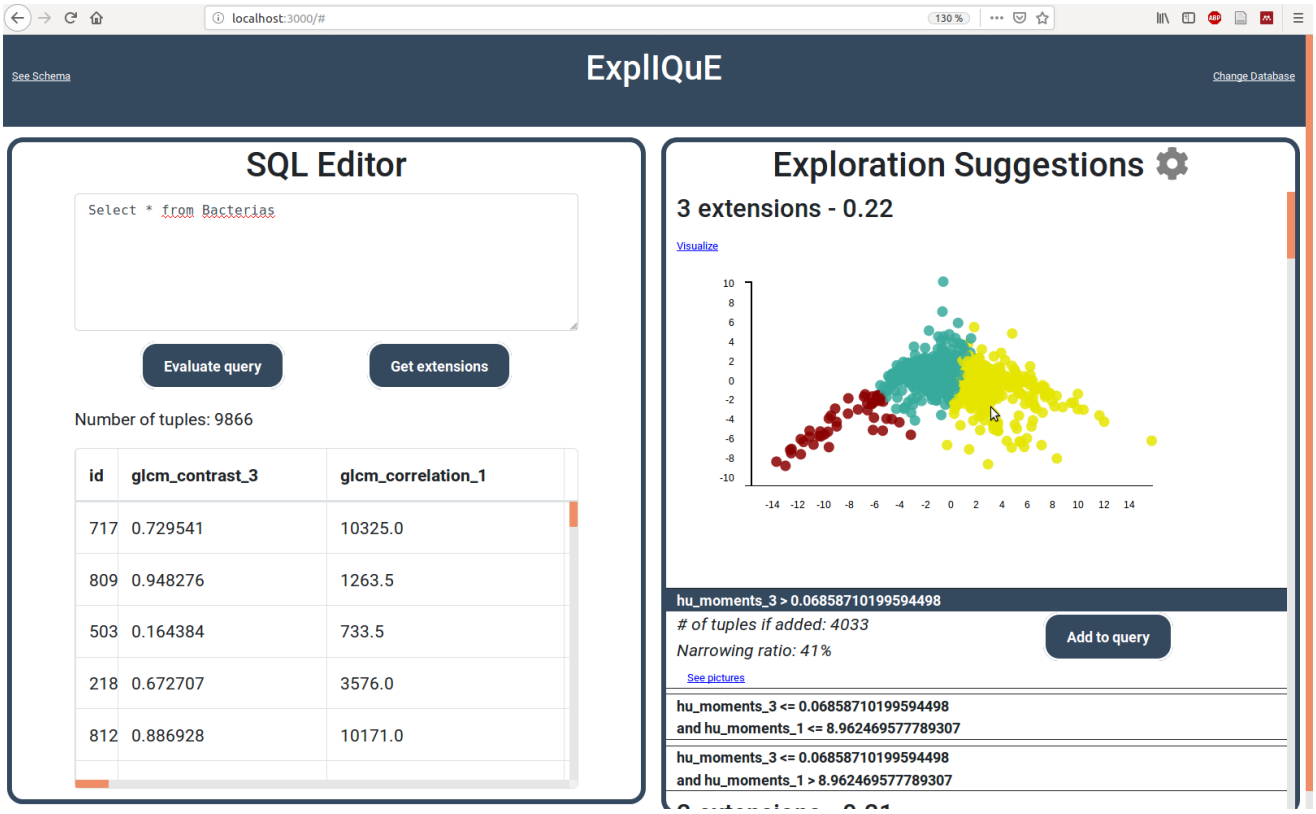


**Figure 1: Web interface for ExplIQuE**

depuis sa racine jusqu'à la feuille, pour agréger les différentes décisions qui mènent jusqu'à la feuille.

L'interaction entre l'utilisateur et les extensions se fait via une interface web, dont une capture d'écran est présentée sur la figure 1. La partie gauche de l'interface est dédiée aux fonctionnalités SQL classiques (saisie de requête et affichage des résultats), la droite à l'affichage des extensions. Lorsque plusieurs valeurs de $k$ sont testées, les différents groupes d'extensions sont classées selon le coefficient de silhouette [4] : cela permet de proposer des solutions même quand l'utilisateur n'a aucune idée du nombre d'extensions désirées. Dans un même groupe, les extensions sont classée de la plus générale (qui contient le plus de tuples) à la plus spécifique. Enfin, une visualisation permet de voir la répartition des tuples initiaux dans chaque extension, en projettant les tuples en 2D en utilisant une analyse en composants principaux (ACP), et en représentant chaque extension par une couleur différente.

## 3 SCÉNARIO DE LA DÉMONSTRATION

L'objectif de la démonstration est de permettre à l'audience de tester les différentes fonctionnalités d'ExplIQuE, en répondant à des questions imprécises sur une base de données scientifiques. Cette base de données vient d'une étude sur des colonies de bactéries [1], et contient 10000 tuples sur 29 colonnes. Ces données décrivent la forme, texture et couleurs de bactéries de différentes colonies. Ainsi, le problème ne se pose que sur une seule table, mais le contenu et le nombre de colonne n'est pas évident à appréhender au premier abord.

Pour la démonstration, nous proposons des questions imrpécises sur ces données, comme par exemple : Quelles sont les bactéries qui ont toutes une texture similaire, et une forme circulaire ? Si cette question peut décontenancer au premier abord, ExplIQuE permet normalement d'y répondre en maximum trois itérations, soit en moins de cinq minutes. Les utilisateurs pourront ainsi jouer sur les paramètres proposés par l'interface pour obtenir le résultat désiré, et explorer les différentes visualisations à leur disposition.

1. Merci à Christopher Pease de Darlington EURL pour le partage de ce jeu de données

## RÉFÉRENCES

[1] Marie Le Guilly, Ihab Ilyas, Jean-Marc Petit, and Vasile-Marian Scuturici. 2018. Partitioning queries for data exploration using query extensions. In *BDA 2018 34ème conférence sur la Gestion de Données. Principes, Technologies et Applications.*

[2] S. Lloyd. 2006. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theor.* 28, 2 (Sept. 2006), 129–137. https://doi.org/10.1109/TIT.1982.1056489

[3] Chaitanya Mishra and Nick Koudas. 2009. Interactive query refinement. In *Proceedings of the 12th International Conference on Extending Database Technology : Advances in Database Technology.* ACM, 862–873.

[4] Peter J Rousseeuw. 1987. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

# Interroger des Lacs de Données en utilisant Spark & Presto

Mohamed Nadjib Mami[†,§], Damien Graux[†,◇], Simon Scerri[†,§], Hajira Jabeen[§], and Sören Auer[£*]
† Fraunhofer IAIS ; § University of Bonn
◇ ADAPT Centre, Trinity College Dublin ; £ TIB and L3S Research Center
{mami,scerri,jabeen}@cs.uni-bonn.de;damien.graux@adaptcentre.ie;auer@l3s.de

## ABSTRACT

Squerall est un outil permettant l'interrogation de sources de données hétérogènes à large échelle en utilisant à bon escient des moteurs de traitement dédiés aux larges volumes de données issus de la littérature: Spark et Presto. Les requêtes à destination des lacs de données sont évaluées à la volée, *i.e.*, directement sur les sources originelles sans procéder à de quelconques transformations préalables des données. Nous démontrons la capacité qu'a Squerall à interagir avec cinq sources différentes parmi lesquelles Cassandra et MongoDB. En particulier, nous mettons en évidence que notre outil peut joindre ensemble plusieurs sources en même temps, tout en montrant qu'étendre la couverture à d'autres sources potentielles reste simple. Des interfaces graphiques sont aussi mises à disposition pour (1) construire les requêtes SPARQL et (2) mettre en place les fichiers de configuration nécessaires.

## CCS CONCEPTS

• **Information systems** → **Database query processing**; **Parallel and distributed DBMSs**; **Mediators and data integration**; • **Applied computing** → **Information integration and interoperability**; • **Computing methodologies** → *Knowledge representation and reasoning*.

## 1 INTRODUCTION

During the last four decades, a variety of data storage and management techniques have been developed in both research and industry. Today, we benefit from a multitude of storage solutions, varying in their data model (e.g. tabular, document, graph) or their ability to scale storage and querying. There are dozens of continuously evolving storage and data management solutions. As a result, users can choose a system that suits their individual application needs. However, those systems do not inter-operate, every stored datum is locked in the respective system it is stored in. For example, an e-commerce company might store *product* information in a Cassandra database, *offers* in MongoDB to benefit from its capability to store hierarchical multi-level values, and information about *Producers* obtained from an external source in a relational format. Without transforming and moving the data into a unified (scalable) data management solution, the data can hardly be explored and business insights be extracted using ad hoc uniform querying. We have taken on the mission of bridging this gap and developed Squerall[1] [9]: a software that gives access to heterogeneous data kept in their original forms and sources using Semantic Web techniques to enable uniform querying with SPARQL[2].

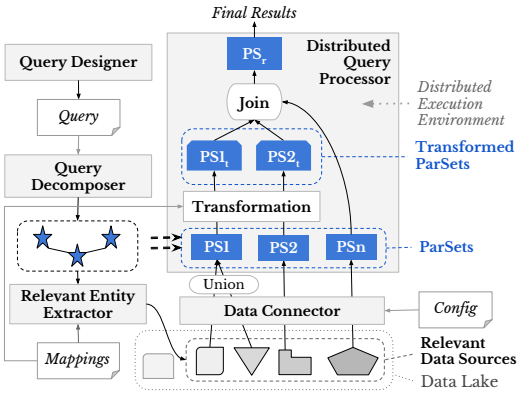


**Figure 1: Squerall High-level Architecture.**

Similar efforts to integrate and query large data sources exist in the literature. For instance, [3] defines a mapping language to express access links to NoSQL databases. [11] allows to run CRUD operations over NoSQL databases. [1] proposes a unifying *programming* model to directly access databases using *get*, *put* and *delete* primitives. [7] proposes a SQL-like language containing invocations to the native query interface of relational and NoSQL databases. [6] is a hybrid platform with consideration for both heterogeneous and dynamic data sources (streams). However, Squerall offers the highest number of supported data sources (namely: CSV, Parquet, Cassandra, MongoDB and MySQL) while providing the richest query capability, including joining, aggregation and ordering.

## 2 SQUERALL: CONCEPTS & ARCHITECTURE

Squerall [9] implements the so-called Ontology-Based Data Access (OBDA) [10] paradigm. In OBDA, data schemata are mapped to higher-level ontologies, forming a middleware against which queries are posed. These SPARQL queries are then executed in a separate distributed *environment*, which is, in particular, resilient to faults (node failure does not halt the entire query execution), and elastic and horizontally-scalable (more nodes can be added to accommodate more expensive computations). In addition, as data from different sources is generated by different applications, they

*This research was partially supported by the European Union's H2020 research and innovation programme BETTER (GA. 776280); and by the ADAPT Centre for Digital Content Technology funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.



[1] Associated website: <https://eis-bonn.github.io/Squerall/>
[2] <http://www.w3.org/TR/sparql11-overview/>

| | SPARQL SELECT Queries adapted from BSBM | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q10 |
| – Scale 1: scale factor of 0.5 millions products – | | | | | | | | | |
| Presto | 55.34 | 28.89 | 15.84 | 53.63 | 49.24 | 43.38 | 18.63 | 14.38 | 89.08 |
| Spark | 98.78 | 189.57 | 59.96 | 277.30 | 222.76 | 191.26 | 159.51 | 91.38 | 300.38 |
| Diff. % | **178.48** | **656.17** | **378.57** | **517.06** | **452.40** | **440.88** | **856.31** | **635.33** | **337.21** |
| – Scale 2: scale factor of 1.5 millions products – | | | | | | | | | |
| Presto | 139.15 | 48.59 | 16.30 | 133.45 | 115.35 | 116.08 | 42.16 | 14.37 | 405.84 |
| Spark | 102.86 | 584.67 | 70.76 | 673.12 | 637.18 | 611.65 | 447.27 | 75.19 | 888.98 |
| Diff. % | 73.92 | **1203.36** | **434.05** | **504.41** | **552.40** | **526.93** | **1060.83** | **523.31** | **219.05** |
| – Scale 3: scale factor of 5 millions products – | | | | | | | | | |
| Presto | 276.91 | 131.87 | 30.58 | 340.61 | 350.04 | 334.29 | 98.11 | 18.96 | 784.01 |
| Spark | 132.37 | 1813.69 | 93.19 | 2131.10 | 1846.59 | 1833.47 | 1390.99 | 79.33 | 2703.43 |
| Diff. % | 47.80 | **1375.40** | **304.71** | **625.67** | **527.54** | **548.46** | **1417.80** | **418.47** | **344.82** |

**Table 1: Query execution times (seconds) using Presto and Spark and the difference percentage between them (%).**

may not be able to be readily cross-joined. Thus, modifications on the possible join values ought to be incorporated. Squerall is comprised of five components (see Figure 1):

- **Query Decomposer:** Validates and analyzes SPARQL queries provided by a user. Particularly, the Query Decomposer extracts the Basic Graph Pattern fragment of the query and decomposes it into star-shaped sub-graph patterns having the same subject, *stars* for short. This component also detects links between stars.
- **Relevant Entity Extractor:** Each star is analyzed separately; this component searches in the mappings for entities that are mapped to every predicate of the star.
- **Data Connector:** Once relevant data entities are detected, they are connected to the distributed execution environment. Every detected entity is loaded into a *ParSet* (Parallel dataSet): a data structure that can be distributed and operated on concurrently. The Data Connector expects users to input connection metadata.
- **Distributed Query Processor:** Following the principles introduced earlier, queries are executed in parallel. Query execution occurs on and across the ParSets. Links between stars retrieved by the Query Decomposer are transformed into joins between the relevant detected data entities and all stars are incrementally joined. When disjoinability points are known, join values are altered to enable joinability.
- **Query Designer:** We see the necessity of supporting users in their SPARQL query creation.

## 3 TECHNICAL DETAILS

**Core technologies.** RML [5] and FNO [4] are used to express mappings and to declare query-independent transformations. *Apache Spark* and *Presto* are used to implement both the Data Connector and Distributed Query Processor. Spark is a general-purpose processing engine, and Presto is a distributed SQL query engine. Both provide dozens of wrappers[3] to connect to a data source. They load the data (fully, or partially) to their in-memory data structures.

**Achieved Performance.** We evaluate the performance querying five different data sources: Cassandra, MySQL, MongoDB, Parquet, and CSV. As evaluation data we choose the BSBM [2] benchmark. We pick five relational tables (*Product, Producer, Offer, Review, and Person*) and load them into the five data sources. For performance and scalability we evaluate the execution time of BSBM queries against three increasing data sizes: 0.5m, 1,5m and 5m scale factors (number of products). We experiment with all BSBM SELECT queries with some adaptation (queries are altered to involve the five tables we populated). The evaluation was carried out in a cluster of three nodes, each having a 16-core DELL PowerEdge processor, 256GB RAM, and 3TB SATA disk. Presto-based Squerall exhibits better performance to the Spark-based alternative in the majority of the cases. To measure this difference, we calculate the difference, in percentage, between Presto and Spark execution times (third row under each scale result in Table 1), e.g., for Q2 in the first scale Presto-based Squerall is 656% faster.

**User Interfaces.** We provide 3 GUIs to help users produce Squerall's inputs: *Config*, *Mappings* files and *SPARQL query*. They have built-in search functionalities that send requests to the LOV catalog[4] to search for adequate terms from existing ontologies.

## 4 CONCLUSION

Squerall addresses the *Variety* challenge of Big Data by making use of Semantic Web standards and best practices. It can be extended to embrace new data sources, by making use of the query engines' own wrappers. Additionally, Squerall has been integrated[5] into SANSA [8], a framework for scalable processing and analysis of large-scale RDF data, widening its scope to also access non-RDF data sources. Squerall source code is available under an Apache-2.0 license on GitHub[6]. In addition, a screencast presenting the various interfaces and the query execution is available[7]. The deployment is further facilitated with a Dockerfile to quickly run the BSBM use-case described here.

## REFERENCES


[1] Paolo Atzeni, Francesca Bugiotti, and Luca Rossi. 2012. Uniform access to non-relational database systems: The SOS platform. In *International Conference on Advanced Information Systems Engineering*. Springer, 160–174.

[2] Christian Bizer and Andreas Schultz. 2009. The berlin SPARQL benchmark. *International Journal on Semantic Web and Information Systems* 5, 2 (2009), 1–24.

[3] Olivier Curé, Robin Hecht, Chan Le Duc, and Myriam Lamolle. 2011. Data integration over nosql stores using access path based mappings. In *International Conference on Database and Expert Systems Applications*. Springer, 481–495.

[4] Ben De Meester, Wouter Maroy, Anastasia Dimou, Ruben Verborgh, and Erik Mannens. 2017. Declarative data transformations for Linked Data generation: the case of DBpedia. In *European Semantic Web Conference*. Springer, 33–48.

[5] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. 2014. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data.. In *LDOW*.

[6] Martin Giese, Ahmet Soylu, Guillermo Vega-Gorgojo, Arild Waaler, Peter Haase, Ernesto Jiménez-Ruiz, Davide Lanti, Martín Rezk, Guohui Xiao, Özgür Özçep, et al. 2015. Optique: Zooming in on big data. *Computer* 48, 3 (2015), 60–67.

[7] Boyan Kolev, Patrick Valduriez, Carlyna Bondiombouy, Ricardo Jimenez-Peris, Raquel Pau, and José Pereira. 2016. CloudMdsQL: querying heterogeneous cloud data stores with a common language. *Distributed and parallel databases* 34, 4 (2016), 463–503.

[8] Jens Lehmann, Gezim Sejdiu, Lorenz Bühmann, Patrick Westphal, Claus Stadler, Ivan Ermilov, Simon Bin, Nilesh Chakraborty, Muhammad Saleem, Axel-Cyrille Ngonga Ngomo, and Hajira Jabeen. 2017. Distributed Semantic Analytics using the SANSA Stack. In *ISWC Resources Track*.

[9] Mohamed Nadjib Mami, Damien Graux, Simon Scerri, Hajira Jabeen, Sören Auer, and Jens Lehmann. 2019. Squerall: Virtual ontology-based access to heterogeneous and large data sources. In *International Semantic Web Conference*. Springer, 229–245.

[10] Antonella Poggi, Domenico Lembo, Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, and Riccardo Rosati. 2008. Linking data to ontologies. In *Journal on Data Semantics X*. Springer, 133–173.

[11] Rami Sellami and Bruno Defude. 2018. Complex queries optimization and evaluation over relational and NoSQL data stores in cloud environments. *IEEE Transactions on Big Data* 4, 2 (2018), 217–230.


[3] <https://spark-packages.org> and <https://prestodb.io/docs/current/connector.html>

[4] Linked Open Vocabularies: publish and search ontologies <https://lov.linkeddata.es/>

[5] <https://github.com/SANSA-Stack/SANSA-DataLake>

[6] <https://github.com/EIS-Bonn/Squerall>

[7] <https://github.com/EIS-Bonn/Squerall/blob/master/evaluation/screencasts>

# CATI: Assisted Classification of Documents (text and images)

Gabriela Bosetti
Előd Egyed-Zsigmond
Lucas Okumura-Ono
gabriela.bosetti@insa-lyon.fr
elod.egyed-zsigmond@insa-lyon.fr
lucas.okumura--ono@insa-lyon.fr
Université de Lyon. UMR 5205 CNRS
Villeurbanne, France

## ABSTRACT

Domain knowledge is essential to Data science since it provides the scope for the construction of models, methods, and techniques that allow extracting insights from large amounts of data. A well-known problem in this multi-disciplinary field is that the person having such an initial knowledge is not necessarily a data scientist but a domain expert, therefore s/he misses the education and experience on data analysis. In this article, we present a full platform for as-sisted classification in the domain of microblogs, mainly conducted by text- and image-based event detection and Active Learning (AL). The process is fully supported through a graphical user interface whose source code is freely accessible, and provides users with classification and data exploration features.

## CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*; *Retrieval tasks and goals*; • **Human-centered computing**;

## KEYWORDS

information retrieval, assisted document classification, active learn-ing, human-computer interaction

**ACM Reference Format:**
Gabriela Bosetti, Előd Egyed-Zsigmond, and Lucas Okumura-Ono. 2019. CATI: Assisted Classification of Documents (text and images). In *Proceed-ings of BDA '19: 35ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA '19).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn


*BDA '19, October 15–18, 2019, Lyon, France*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

# Exploring and Curating Data Collections with CURARE

Genoveva Vargas-Solar
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG-LAFMIA
Grenoble, France
genoveva.vargas@imag.fr

Gavin Kemp
Univ Lyon, University of Lyon 1, LIRIS UMR 5205 CNRS
Villeurbanne, France
gavin.kemp@liris.cnrs.fr

Irving Hernández-Gallegos
Universidad Autónoma de Guadalajara
Zapopan, Mexico
irving.hernandez.g@gmail.com

Javier A. Espinosa-Oviedo
Delft University of Technology
2628BL Delft, Netherlands
javier.espinosa@tudelft.nl

Catarina Ferreira Da Silva
Univ Lyon, University of Lyon 1, LIRIS UMR 5205 CNRS
Villeurbanne, France
catarina.ferreira-da-silva@liris.cnrs.fr

Parisa Ghodous
Univ Lyon, University of Lyon 1, LIRIS UMR 5205 CNRS
Villeurbanne, France
parisa.ghodous@liris.cnrs.fr

**ABSTRACT**

This paper demonstrates CURARE, an environment for curating and assisting data scientists to explore *raw* data collections. CU-RARE implements a data curation model used to store structural and quantitative metadata semi-automatically extracted. It provides associated functions for exploring these metadata. The demonstra-tion proposed in this paper is devoted to evaluate and compare the effort invested by a data scientist when exploring data collections with and without CURARE assistance.

Demonstration already presented and published in the conference EDBT 2019.


ACM ISBN 978-3-89318-081-3.

# BeLink: Querying Networks of Facts, Statements and Beliefs

Tien-Duc Cao
Inria and LIX (UMR 7161, CNRS and Ecole polytechnique)
Palaiseau, France
tien-duc.cao@inria.fr

Ludivine Duroyon
Univ Rennes, CNRS, Inria, IRISA
Lanion, France
ludivine.duroyon@irisa.fr

François Goasdoué
Univ Rennes, CNRS, Inria, IRISA
Lanion, France
fg@irisa.fr

Ioana Manolescu
Inria and LIX (UMR 7161, CNRS and Ecole polytechnique)
Palaiseau, France
ioana.manolescu@inria.fr

Xavier Tannier
Sorbonne Université and Inria
Paris, France
xavier.tannier@sorbonne-universite.fr

**ABSTRACT**

An important class of journalistic fact-checking scenarios [2] in-volves verifying the *claims* and *knowledge* of different actors at different moments in time. Claims may be about *facts*, or about other claims, leading to chains of hearsay. We have recently pro-posed [4] a data model for (time-anchored) facts, statements and beliefs. It builds upon the W3C's RDF standard for Linked Open Data to describe connections between agents and their statements, and to trace information propagation as agents communicate. We propose to demonstrate BeLink, a prototype capable of storing such interconnected corpora, and answer powerful queries over them relying on SPARQL 1.1. The demo will showcase 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.

# 8 Prix du meilleur article.

### Comité du prix du meilleur article.

**Karine Zeitouni**
DAVID - Université de Versailles Saint-Quentin

- Nicole Bidoit-Tollu (LRI- Univ. Paris-Sud)
- David Gross-Amblard (IRISA - Univ. Rennes 1)
- Zoubida Kedad (DAVID, Univ. Versailles Saint-Quentin)
- Laurent d'Orazio (IRISA - Univ. Rennes 1)
- Genoveva Vargas-Solar (LIG & LAFMIA - CNRS)
- Dan Vodislav (ETIS - Univ. Cergy-Pontoise)

### Lauréats du prix du meilleur article.

Prix du Meilleur Article **Ousmane Issa**, **Angela Bonifati** et **Farouk Toumani** pour l'article « *A Relational Framework for Inconsistency-aware Query Answering* »

Prix du Meilleur Article **Thomas Minier**, **Hala Skaf-Molli** et **Pascal Molli** pour l'article « *SaGe : Web Preemption for Public SPARQL Query Services* »

# 9 Prix de la meilleure démonstration.

### Comité du prix de la meilleure démonstration.

**Yehia Taher**
DAVID - Université de Versailles Saint-Quentin

Comité ad-hoc.

### Lauréats du prix de la meilleure démosntration.

Prix de la meilleure démonstration **Yanlei Diao**, **Pawel Guzewicz**, **Ioana Manolescu** et **Mirjana Mazuran** pour la démonstration intitullée :
« *Spade : a Modular Framework for Analytical Exploration of RDF Graphs* »

# 10 Prix de thèse en gestion de données.

Pour la première fois en 2019, la communauté BDA distingue par un prix (et éventuellement un accessit) une thèse réalisée dans le domaine de la gestion de données (au sens large). Sont éligibles les doctorant(e)s :

— ayant publié au moins une fois à la conférence BDA (y compris en 2019), un article long, un article doctorant, et/ou une démonstration ;

— ayant soutenu (ou prévoyant de soutenir) leur thèse dans l'année calendaire qui finit avant la conférence BDA 2019. Exceptionnellement pour cette 1ère année, nous prendrons en considération un intervalle légèrement plus grand, allant du 01/09/2018 au 14/10/2019.

— disposant au 15 juillet 2019 de la version du manuscrit envoyée aux rapporteurs, ainsi que du formulaire de composition du jury ; ces deux éléments devront être déposés sur le site de candidature. De plus, pour que la thèse soit considérée pour le prix, il faut transmettre au jury avant le 14/10/2019 les rapports sur le manuscrit et le rapport de soutenance.

Le prix distingue les contributions apportées pendant la thèse, que ce soit au niveau formel, théorique, d'architecture, et/ou de développement d'algorithmes, prototypes ou systèmes. Sont éligibles toutes les thèses soutenues dans les conditions ci-dessus, dans les domaines scientifiques concernés par les appels à communications récents des conférences BDA (2019, 2018, 2017, 2016).

### Comité du prix de la meilleure thèse en gestion de données.

**Ioana Manolescu**
Inria Saclay

- Sihem Amer-Yahia (LIG - CNRS, Grenoble)
- Nicolas Anciaux (Inria & DAVID - Univ. Versailles Saint-Quentin)
- Dario Colazzo (LAMSADE - Univ. Paris-Dauphine)
- Florent Masseglia (Inria & LIRMM, Univ. Montpellier)
- Jean-Marc Petit (LIRIS - INSA-Lyon)
- Pierre Senellart (Inria & DIENS, ENS Paris)
- Farouk Toumani (LIMOS - Univ. Auvergne)

### Lauréats du prix de la meilleure thèse en gestion de données.

Prix de thèse **Michele Linardi** pour sa thèse intitulée
« *Variable-length Similarity Search for Very Large Data Series : Subsequence matching, Motif and Discord Detection* ».

Prix de Thèse **Mikaël Monet** pour sa thèse intitulée
« *Combined Complexity of Probabilistic Query Evaluation* »

Accessit au Prix de Thèse **Ugo Comignani** pour sa thèse intitulée
« *Interactive Mapping Specification and Repairing in the Presence of Policy Views* »