Activity Report 2022

Team DRUID

Declarative & Reliable management of Uncertain user-generated Interlinked Data

D7 – Data and Knowledge Management

1 Team composition

Researchers and faculty

Jean-Christophe Dubois, Associate Professor, IUT Lannion - Lannion
Mickaël Foursov, Associate Professor, ISTIC Univ. Rennes - Rennes
David Gross-Amblard, Professor, ISTIC Univ. Rennes - Rennes
Yolande Le Gall, Associate Professor, IUT Lannion, Univ. Rennes - Lannion
Arnaud Martin, Professor, IUT Lannion, Univ. Rennes - Lannion
Zoltan Miklos, Associate Professor, ESIR Univ. Rennes - Rennes, **head of the team**
Virginie Sans, Associate Professor, ISTIC Univ. Rennes - Rennes
Constance Thierry, Associate Professor (from 09/22), ATER until 08/2022, IUT Lannion, Univ. Rennes - Lannion

Engineer

Maria Massri (11/2022-08/2023)

ATER teaching/research assistantship

Mathieu Chambe (09/2022-08/2023)

PhD students

Maria Massri, CIFRE OrangeLabs (until December 2022)
Fancois Mentec, CIFRE ALTEN
Arthur Hoarau, Departemental and regional financement
Mathieu Chambe, bourse (ex-team PERCEPT)
Erwan Vincent, CIFRE KEOLIS
Aymen Bazouzi, CominLabs Clara project
Zuowei Zhang, CSC grant, (until January 2022)

Administrative assistant

Gunther Tessier (INRIA)

2 Overall objectives

2.1 Overview

Recently, there is an increased interest in data management methods. Statistical machine learning techniques, empowered by the available pay-as-you-go distributed computing power, are able to extract useful information from certain data. The international press, being specialized or not, has echoed these remarkable results as a new Spring for Artificial Intelligence in a broad sense. The data is sometimes even referred to as the "gold of the 21st century'. In any area of business and science, one tries to construct huge datasets to be able to profit from the benefits of the Artificial intelligence revolution.

Our team works on questions of data management techniques to efficiently store, query and organise data. We also work on artificial intelligence techniques to extract knowledge, and to gain understanding from data, especially in the presence of uncertainties. Ideally this knowledge should be actionable to be able to provide services based on them (e.g. recommendations).

Unfortunately, data management and machine learning are often seen as different tasks. Machine learning primitives are not supported elegantly for now, in the data management dogma. For example, Machine Learning operators are seen for now as external procedures outside the query language, barely accounted for by the optimizer. Moreover, the knowledge extraction tasks are hard to design without understanding the available data, thus one should consider knowledge extraction as an interactive process, where users influence the process.

The above listed observations lead us to define the following goals for the DRUID team:

- Propose new query mechanisms, in particular for network oriented data and to better integrate Machine Learning methods with the database logic and engines

- Propose interactive, human-in-the-loop data analysis and knowledge extraction methods even with uncertain data

2.2 Scientific foundations

Our team gathers specialists from data management, information extraction and belief functions, various bricks that contribute to our goal. As a common ground, for data management we will naturally elaborate on classical techniques: finite model theory, complexity theory, declarative or algebraic languages, execution plans, costs models, storage and indexing strategies. The theory of belief functions (also commonly referred to as Dempster-Shafer theory) allows to take simultaneously into account both uncertainty and imprecision on the data but also on the models. This theory is one of the most popular one among the quantitative approaches because it can be seen as a generalization of both classical probabilities and possibilities theories. Belief functions are especially developed for information fusion, pattern recognition and clustering.

Analytics in databases Making sense of large amounts of data and extracting useful information is a problem in various fields, in business context as well in various scientific domains. One needs to rely on a wide range of techniques (regression, clustering, embeddings, . . .). A classical data analytics workflow is 1) to extract, to model imperfection and clean a data set, 2) to learn a model and to consider imperfection and 3) to make predictions. Such workflows are now very well handled in procedural languages such as Python or Scala, at various scales (*e.g.* Big Data in Spark).

While this approach works well, it it does not make use the numerous achievements in the database field: when the data set is updated, the workflow has to be re-run (dynamicity problem), data are now much more evolved than classical numerical or categorical ones, such as graphs (data type problem), and machine learning operators are not first class citizen in database query languages (closure problem). Moreover, it is often impossible to formulate the "right" knowledge extraction or machine learning tasks, as it would require the knowledge of the large and heterogeneous datasets, *a priori*.

Our specific goal is to develop data management and -possibly interactive- data analysis methods for generic, uncertain and time-varying data (*e.g.* large evolving graphs). We will rely on graph signal processing [SNF+13], spectral graph theory [Chu97], graph neural networks [WPC+20], graph databases [RWE15], [BFVY18] and graph embedding techniques. We aim at modelling and querying graphs, with (temporal) integrity constraints, where graph analytics is first used to optimize data storage and evaluation. Machine learning techniques also allow to build realistic huge benchmark data sets, that do not exist for all domains.

[BN03]

On a longer perspective, we would like to work on other aspects of database and machine learning integration. In particular, databases have efficient mechanisms for indexing and loading data to main memory and these could be better exploited to realize machine learning tasks. In some cases one could envisage that the machine learning tasks are realized inside the database systems and machine learning methods use database primitives [GR17]. Other potential direction is to consider a vector-space embedding of entire relational databases [Gro20] that could open entire new ways to

[SNF+13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains", *IEEE Signal Processing Magazine 30*, 3, May 2013, p. 83–98.

[Chu97] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[WPC+20] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P. S. Yu, "A Comprehensive Survey on Graph Neural Networks", *IEEE Transactions on Neural Networks and Learning Systems*, 2020, p. 1–21.

[RWE15] I. Robinson, J. Webber, E. Eifrem, *Graph Databases: New Opportunities for Connected Data*, edition 2nd, O'Reilly Media, Inc., 2015.

[BFVY18] A. Bonifati, G. H. L. Fletcher, H. Voigt, N. Yakovets, *Querying Graphs*, *Synthesis Lectures on Data Management*, Morgan & Claypool Publishers, 2018, https://doi.org/10.2200/S00873ED1V01Y201808DTM051.

[BN03] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", *Neural Comput. 15*, 6, June 2003, p. 1373–1396, https://doi.org/10.1162/089976603321780317.

[GR17] M. Grohe, M. Ritzert, "Learning first-order definable concepts over structures of small analyze data stored in such systems.

2.3 Application domains

Our natural applications are storing and querying large-scale semantic graphs for IOT (*e.g.* Maria Massri's thesis), Digital humanities (epistemology, understanding the evolution of ideas and scientific fields, (*e.g.* EPIQUE ANR project, completed), human resources management (*e.g.* François Mentec's thesis), and crowd management systems (HEADWORK ANR) for "artificial artificial intelligence". Our work and results can be used to analyze the evolution of other types of networks (e.g. transportation network, e.g. Erwan Vincent, Gauthier Lyan's thesis) and in the areas of IA and education (Clara project).

3 Scientific achievements

3.1 Temporal graph databases

Participants: Zoltan Miklos, David Gross-Amblard, Maria Massri.

Our work is focused on questions related to temporal graph databases. Specifically, we further improved our δ-Copy+Log storage method, our space-efficient and configurable storage technique and our article was finally accepted at ICDE'2022 [13], one of the leading conferences in data management.

We also worked on temporal graph generation. The goal of this work is to generate large temporal graphs that one can use to evaluate temporal graph databases. Our generation techniques enable to generate graph with a given degree distribution, and also with a give community structure. Our assumption -motivated by our IoT use case- is that the large graphs evolve gradually, and there is only change between consecutive snapshots. We obtain the temporal graph with the help of an optimal transport solver. We have developed a tool that realizes our generation methods. Our work was published at the DataPlat workshop, at EDBT'2022 [14]. We have prepared a more complete version of our article and submitted to a journal. The article is under review (we submitted the revised manuscript in December 2022).

Furthermore we forked on query evaluation methods, in the specific context of temporal graph databases. We have various publications in preparation.

Maria Massri defended her thesis in December 2022. Her thesis will be available in HAL shortly.

degree", *in: 32nd Annual ACM/IEEE Symposium on Logic in Computer Science, LICS 2017, Reykjavik, Iceland, June 20-23, 2017*, IEEE Computer Society, p. 1–12, 2017, `https://doi.org/10.1109/LICS.2017.8005080`.

[Gro20] M. Grohe, "word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data", *CoRR abs/2003.12590*, 2020, `https://arxiv.org/abs/2003.12590`.

3.2 AI for human resource applications

Participants: Zoltan Miklos, Francois Mentec.

We have developed a recommender system that can support HR experts in the recruitment process. Specifically the tool is used on a daily basis for internal recruitment at the consulting company ALTEN. The recommendations are generated with the help of word embedding (BERT) techniques. We have prepared a publication that presents the recommendation system. We evaluated the improvement of the recruitment process through this recommendation service. Unfortunately the paper was not accepted, but we are about the resubmit the work at another venue. The recommendation service is used in a daily basis at the consulting company ALTEN: this project is a clear success from the point of view of industrial adoption.

3.3 AI for education

Participants: Zoltan Miklos, Mickael Foursov, David Gross-Amblard, Aymen Bazouzi. Collaboration with Hoel Le Capitaine (Nantes University), Annie Foret (ex-Semlis team), Sebastian Ferre (Lacodam).

We have launched our work on AI and education in the context of Clara project (comin-labs) in 2022. The project addresses questions about the combination of open educational resources. Our main activity is the development of a recommendation service for educational resources, with the help of graph representation learning techniques. We are about to prepare a publication on this subject, specifically we developed a representation learning technique that exploits a graph structure that one can associate to the concepts of a learning resource. We will continue our work on this topic and exploit this representation to realise classification or recommendation services that could be used for open educational resources.

Another important question is the formal representation of licences that are often expressed in natural language and the reasoning about usage rights of resources, based on their licences. We collaborate with Annie-Foret on this subject and develop reasoning techniques, based on graph rewriting methods. In this context we will host an intern (Malo Revel, ENS Rennes, M2 SIF) (supervisors Annie Foret, and Zoltan Miklos).

Furthermore the Clara project is likely to rely on crowdsourcing services to annotate learning resources or evaluate the quality of certain courses or course materials.

3.4 AI for public transport simulation and understanding

Participants: Zoltan Miklos, Simon Malinowski (linkmedia), Guillaume Gravier (linkmedia).

We started our collaboration with KEOLIS in 2022, Erwan Vincent started his thesis in February 2022. However, the DRUID team was engaged already in a collaboration with KEOLIS before (supervision of the thesis of Gauthier Lyan). In 2022, Erwan revitalised the work of Gauthier and significantly improved the predictive models, developed by

Guthier. The work involved large and heterogeneous data collection and the development of more complete predictive models. We are about to prepare a publication on the improvements of the bus punctuality predictions. We plan to further improve these predictions, specifically to develop models that exploit data from the entire public transport network, rather to make local predictions. We are also investigating how to simulate the public transport network to support the development of the high-quality bus service that the company is about to prepare (to launch in 2030).

3.5 AI and belief functions

Participants: Arnaud Martin, Arthur Hoarau, Zuowei Zhang, Yiru Zhang, Constance Thierry, Yolande Le Gall, Jean-Christophe Dubois.

3.5.1 Belief clustering

Participants: Arnaud Martin, Zuowei Zhang, Yiru Zhang.

Clustering is an essential part of data mining, which can be used to organize data into sensible groups. We continue to develop some research to propose new clustering method adapted to imperfect data.

The theory of belief functions (TBF) is now a widespread framework to deal and reason with uncertain and imprecise information, in particular to solve information fusion and clustering problems. Combination functions (rules) and distances are essential tools common to both the clustering and information fusion problems in the context of TBF, which have generated considerable literature. Distances and combination between evidence corpus of TBF are indeed often used within various clustering and classification algorithms, however their interplay and connections have seldom been investigated, which is the topic of this paper. More precisely, in [8], we focus on the problem of aggregating evidence corpus to obtain a representative one, and we show through an impossibility theorem that in this case, there is a fundamental contradiction between the use of conjunctive combination rules on the one hand, and the use of distances on the other hand. Rather than adding new methodologies, such results are instrumental in guiding the user among the many methodologies that already exist. To illustrate the interest of our results, we discuss different cases where they are at play. Within the theory of belief functions, both distances and conjunctive combination rules can be used to achieve very similar purposes: evaluating the conflict between sources, performing supervised or unsupervised learning in presence of evidential information, or more simply obtaining a synthetic representation of multiple items of information. However, the results obtained by both approaches may show some inconsistency between them. This paper provides some insight as to why this may happen, showing that the two approaches are definitely at odds, and that using distances is, for instance, incompatible with some fundamental notions of the theory of belief functions, such as the least commitment principle. We illustrate the importance of the studied differences on problems such as k-centroid clustering, and discuss the importance of interpretations in such problems, which is rarely done in the literature.

It is still a challenging problem to characterize uncertainty and imprecision between specific (singleton) clusters with arbitrary shapes and sizes. In order to solve such a problem, we propose in [9] a belief shift clustering (BSC) method for dealing with object data. The BSC method is considered as the evidential version of mean shift or mode seeking under the theory of belief functions. First, a new notion, called belief shift, is provided to preliminary assign each query object as the noise, precise, or imprecise one. Second, a new evidential clustering rule is designed to partial credal redistribution for each imprecise object. To avoid the âuniform effectâ and useless calculations, a specific dynamic framework with simulated cluster centers is established to reassign each imprecise object to a singleton cluster or related meta-cluster. Once an object is assigned to a meta-cluster, this object may be in the overlapping or intermediate areas of different singleton clusters. Consequently, the BSC can reasonably characterize the uncertainty and imprecision between singleton clusters. The effectiveness has been verified on several artificial, natural, and image segmentation/classification datasets by comparison with other related methods.

3.5.2 Belief learning

Participants: Arnaud Martin, Zuowei Zhang, Arthur Hoarau, Yolande Le Gall, Jean-Christophe Dubois.

The classification analysis of missing data is still a challenging task since the training patterns may be insufficient and incomplete in many fields. To train a high-performance classifier and pursue high accuracy, in [10] we learn a credal classifier based on an optimized and adaptive multi-estimation (OAME) method for missing data imputation on training and test sets. In OAME, some incomplete training patterns are estimated as multiple versions by a global optimization method thereby expanding the training set. On the other hand, the test pattern is adaptively estimated as one or multiple versions depending on the neighbors. For the test pattern with multiple versions, the corresponding outputs with different discounting factors (weights), represented by the basic belief assignments (BBAs), are fused for final credal classification based on evidence theory. The discounting factor contains two aspects: the importance and reliability factors that are used respectively to quantify the importance of the edited version itself and to represent the reliability of the classification result of the version. The effectiveness of OAME is widely validated on several real datasets and critically compared to other related methods.

Classification is used to predict classes by extracting information from labeled data. But sometimes the collected data is imperfect, as in crowdsourcing where users have partial knowledge and may answer with uncertainty or imprecision. In [11] we offer a way to deal with uncertain and imprecise labeled data using Dempster-Shafer theory and active learning. An evidential version of K-NN that classifies a new example by observing its neighbors was earlier introduced. We propose to couple this approach with active learning, where the model uses only a fraction of the labeled data, and to compare it with non-evidential models. A new computable parameter for EK-NN is introduced, allowing the model to be both compatible with imperfectly labeled data and equivalent to its first version in the case of perfectly labeled data. This method increases the complexity but provides a way to work with imperfectly labeled data with efficient results and reduced labeling costs when coupled with active learning. We have conducted tests on real data imperfectly labeled during crowdsourcing campaigns

3.5.3 Belief functions and crowdsourcing

Participants: Arnaud Martin, Constance Thierry, Yolande Le Gall, Jean-Christophe Dubois.

Crowdsourcing is the outsourcing of tasks to a crowd of contributors on dedicated platforms. The tasks are simple and accessible to all, that is why the crowd is made of very diverse profiles, but this induces con result the contributions of unequal quality. The aggregation method most used in platforms does not take into account the imperfections of the data related to human contributions, which impacts the results obtained. Thus, we propose a new interface for crowdsourcing offering more expression capacity to the contributor. The experiments carried out allowed us to highlight a correlation between the difficulty of the task, the certainty of the contributor and the imprecision of his answer. We use this interface in [15] in order to build a datasets. Indeed, few real, imprecise and uncertain datasets exist to test approaches using belief functions. We have built real birds datasets thanks to the collection of numerous human contributions that we make available to the scientific community. The interest of our datasets is that they are made of human contributions, thus the information is therefore naturally uncertain and imprecise. These imperfections are given directly by the persons. This article presents the data and their collection through crowdsourcing and how to obtain belief functions from the data

Most questionnaires in crowdsourcing offer ordered responses whose order is poorly studied via belief functions. In [12], we study the consequences of a frame of discernment consisting of ordered elements on belief functions. This leads us to redefine the power space and the union of ordered elements for the disjunctive combination. We also study distances on ordered elements and their use. In particular, from a membership function, we redefine the cardinality of the intersection of ordered elements, considering them fuzzy.

3.6 Impact of Data Cleansing for Urban Bus Commercial Speed Prediction

Participants: Gauthier Lyan, Jean-Marc Jézéquel, David Gross-Amblard, Simon Malinowski (Diverse and linkMedia teams).

Public Transportation Information Systems (PTIS) are widely used for public bus services amongst cities in the world. These systems gather information about trips, bus stops, bus speeds, ridership, etc. This massive data is an inviting source of information for machine learning predictive tools. However, it most often suffers from quality deficiencies, due to multiple data sets with multiple structures, to different infrastructures using incompatible technologies, to human errors or hardware failures. In this paper, we consider the impact of data cleansing on a classical machine-learning task: predicting urban bus commercial speed. We show that simple, transport specific business and quality rules can drastically enhance data quality, whereas more sophisticated rules may offer little improvements despite a high computational cost [6].

3.7 Reasoning over Time into Models with DataTime

Participants: Gauthier Lyan, Jean-Marc Jézéquel, David Gross-Amblard, Romain Lefeuvre, Benoit Combemale (Diverse team).

Models at runtime have been initially investigated for adaptive systems. Models are used as a reflective layer of the current state of the system to support the implementation of a feedback loop. More recently, models at runtime have also been identified as key for supporting the development of full-fledged digital twins. However, this use of models at runtime raises new challenges, such as the ability to seamlessly interact with the past, present and future states of the system. In this paper, we propose a framework called DataTime to implement models at runtime which capture the state of the system according to the dimensions of both time and space, here modeled as a directed graph where both nodes and edges bear local states (ie. values of properties of interest). DataTime offers a unifying interface to query the past, present and future (predicted) states of the system. This unifying interface provides i) an optimized structure of the time series that capture the past states of the system, possibly evolving over time, ii) the ability to get the last available value provided by the systemâs sensors, and iii) a continuous micro-learning over graph edges of a predictive model to make it possible to query future states, either locally or more globally, thanks to a com- position law. The framework has been developed and evaluated in the context of the Intelligent Public Transportation Systems of the city of Rennes (France). This experimentation has demonstrated how DataTime can be used for managing data from the past, the present and the future, and facilitate the development of digital twins [7].

3.8 Inverse Tone Mapping using FusionNetwork

Participants: Mathieu Chambe, Zoltan Miklos, Ewa Kijak (Linkmedia team), Kadi Bouatouch (emeritius prof).

We develop a deep learning models that can address computer vision tasks, specifically for high dynamic range (HDR) images. Our approach generates HDR images from standard dynamic range (SRD) images, with the help of inverse tone mapping operators. We use then these generated images to train our network. This approach not only enables to train the network with a smaller dataset, but also results a smaller network, while the performance remains comparable to other state-of-the-art approaches. This is a work is submitted and currently under review.

Mathieu Chambe also published a paper based on his earlier results [4]. His is advancing well with his thesis, we estimate that he will defend his thesis at spring 2023.

4 Software development

4.1 HEADWORK platform

Participants: David Gross-Amblard.

Crowdsourcing relies on potentially huge numbers of on-line participants to resolve data acquisition or analysis tasks. It is an exploding area that impacts various domains, ranging from scientific knowledge enrichment to market analysis support. But currently, existing crowd platforms rely mostly on low level programming paradigms, rigid data models and poor participant profiles, which yields severe limitations. The low- level nature of existing solutions prevents the design of complex data acquisition workflows, that could be executed, composed, searched and even be proposed by participants them- selves. Taking into account the quality, uncertainty, inconsistency and representativeness of participant contributions is still an open problem. Methods for assigning a task to the correct participant according to his trust, motivation and expertise, automatically improving crowd execution time, computing optimal participant rewards, are missing. Similarly, usual crowd campaigns produce isolated and rigid data sets: A flexible and common data model for the produced knowledge about data and participants could allow participative knowledge acquisition. To overcome these challenges, Headwork[1] will define:

Rich workflow, participant, data and knowledge models to capture various kind of crowd applications with complex data acquisition tasks and human specificities Methods for deploying, verifying, optimizing, but also monitoring and adapting crowd- based workflow executions at run time.

To reach its goals, Headwork will rely on two experts of large participative knowledge acquisition platforms: Cesco (Museum National d'Histoire Naturelle), Wirk (Foule-Factory), Valda (INRIA Paris), Druid (Rennes 1), Links (Inria-Lille), Sumo (Inria-Bretagne), Spirals (Inria-Lille).

Over the period of this report the platform is live in Beta version, holding several experimental crowd campaigns. The overall project on GitLab has now 781 commits, 21 members, 5500 PHP lines.

5 Contracts and collaborations

5.1 National Initiatives

5.1.1 Project HEADWORK

Participants: Jean-Christophe Dubois, David Gross-Amblard [contact point], Yolande Le Gall, Arnaud Martin, Zoltan Miklos, Constance Thierry.

- Project type: ANR

[1] https://headwork.irisa.fr

- Dates: 2016-2022
- Coordinator: David Gross-Amblard
- Funding: 800 000 euros / 146 000 euros (IRISA)
- PI institution: ANR
- Other partners: SUMO/IRISA, Cristal and Inria (Lille), MNHN (Paris), ENS and Inria (Paris), FouleFactory/Wirk (startup).

Crowdsourcing relies on potentially huge numbers of on-line participants to resolve data acquisition or analysis tasks. It is an exploding area that impacts various domains, ranging from scientific knowledge enrichment to market analysis support. But currently, existing crowd platforms rely mostly on low level programming paradigms, rigid data models and poor participant profiles, which yields severe limitations. The low- level nature of existing solutions prevents the design of complex data acquisition workflows, that could be executed, composed, searched and even be proposed by participants them- selves. Taking into account the quality, uncertainty, inconsistency and representativeness of participant contributions is still an open problem. Methods for assigning a task to the correct participant according to his trust, motivation and expertise, automatically improving crowd execution time, computing optimal participant rewards, are missing. Similarly, usual crowd campaigns produce isolated and rigid data sets: A flexible and common data model for the produced knowledge about data and participants could allow participative knowledge acquisition. To overcome these challenges, Headwork project aims: 1) Rich workflow, participant, data and knowledge models to capture various kind of crowd applications with complex data acquisition tasks and human specificities, 2) Methods for deploying, verifying, optimizing, but also monitoring and adapting crowd- based workflow executions at run time.

5.1.2 Project Clara

Participants: Zoltan Miklos [Contact point], Mickael Foursov, David Gross-Amblard.

- Project type: Cominlabs
- Dates: 2021.12-2024.12
- Coordinator: Patricia Serano (Nantes Université)
- Funding: 113 000 euros (IRISA)
- PI institution: CominLabs
- Other partners: LS2N (Nantes), IRISA (Rennes)

CLARA project aims to empower teachers to facilitate the creation of licensable educational resources based on existing ones. Our approach will suggest a relevant set of educational resources such that these are coherent with a course sketch and have compatible licenses. The main challenges we will face are how to enrich a network of educational resources using AI algorithms, and how to guarantee a minimal set of license-compatible educational resources relevant to a given course goal with query relaxation techniques. We will exploit educational resources provided by the French Ministry of Education and the X5-GON project.

5.2 Bilateral industry grants

5.3 Cifre ALTEN (until July 2023)

We collaborate with the company ALTEN, who finance in the form of a CIFRE contract the PhD thesis of Francois Mentec. Our collaboration focuses on the use of artificial intelligence techniques for recruitment and human resources tasks in general. In our collaborations we try to propose new ways how the artificial intelligence methods can support the work of recruiters. We do not intend to replace human recruiters or automatically affect consultants to project, rather to support the work of RH agents.

This collaboration was prolonged, as the advancement of thesis of Francois Mentec is slower as expected, however the contract will terminate in July 2023. However, we are discussing the preparation of a second CIFRE contract, on a different application domain. Annie Foret and Olivier Ridoux (ex-Semlis team) were involved in the discussions and they will be potentially co-supervisors of this thesis.

5.4 Cifre OrangeLabs (until October 2022)

We collaborate with the company Orange, who finance in the form of a CIFRE contract the PhD thesis of Maria Massri. The company develops a new platform (called ThingIn) for IoT devices. To realize the proposed services they need to store and query temporal graph data. Our collaboration is on the questions of efficient storage and querying techniques for temporal graph-oriented data.

This collaboration contract has ended by the end of October 2022. However, we plan to continue the collaboration on graph databases, and also on the analysis of graph oriented data and we envisage to establish new contracts.

5.5 Cifre KEOLIS (February 2022 - January 2025)

We collaborate with the company Orange, which finances in the form of a CIFRE contract for the Ph.D. thesis of Erwan Vincent. KEOLIS is the company that operates the public transport service of the city of Rennes. The company would like to understand the factors that influences the commercial speed of the bus services, based on collected data. In particular, they would like to understand these questions, to prepare and to provision the Trambus service that they should operate from 2030. We collaborate with the company, to develop suitable prediction and simulation methods. This work is a collaboration with the LINKMEDIA team.

6 Dissemination

6.1 Promoting scientific activities

6.1.1 Scientific Events Organisation

General Chair, Scientific Chair

Member of the Organizing Committees

6.1.2 Scientific Events Selection

Chair of Conference Program Committees

Member of Conference Program Committees

- D. Gross-Amblard: PC member: BDA'2021, HMData'2021
- A. Martin: PC member: NeurIPS 2022, Belief 2022, Fusion 2022, LFA 2022, EGC'2023,
- Z. Miklos: PC member: SIGKDD'2022, CIKM'2022, WSDM'2022, DSAA'2022, EGC'2023, SoICT 2022, BDA'2022

Reviewer

- C. Thierry: FUSION'2022, Belief 2022, EGC'2023

6.1.3 Journal

Member of the Editorial Boards

- Zoltan Miklos: Editor, Transactions on Large-Scale Data- and Knowledge-Centered Systems LI, Special Issue on Data Management - Principles, Technologies and Applications (Springer LNCS, volume 13410) [1]

Reviewer - Reviewing Activities

- A. Martin: PC member: Applied Soft Computing Journal, Chinese Journal of Aeronautics, Data & Knowledge Engineering, Expert Systems With Applications, International Journal of Approximate Reasoning, Information Sciences
- Z. Miklos: Artificial intelligence, Software Quality journal, Transactions on Moblie computing
- C. Thierry: Decision Support Systems

6.1.4 Invited Talks

- A. Martin : "Conflict management in information fusion", Apprentissage automatique multimodal et fusion d'informations (3iÃ¨me Ã©dition), journÃ©e GDR ISIS, Paris, 15 December 2022.

6.1.5 Leadership within the Scientific Community

- Arnaud Martin:
  - president of EGC society[2]
- David Gross-Amblard:
  - member of BDA board[3]
- David Gross-Amblard, Zoltan Miklos
  - In charge of the website of the French research in database community (`https://bdav.irisa.fr/`)

6.1.6 Scientific Expertise

- A. Martin : ANR, ANRT

6.1.7 Research Administration

6.2 Teaching, supervision

6.2.1 Teaching

- Our team is in charge of most of the database-oriented courses at University of Rennes 1 (ISTIC department and ESIR Engineering school), with courses ranging from classical databases to business intelligence, database theory, MapReduce paradigm, or database security and privacy. We also teach several modules related to machine learning or artificial intelligence.
- Master Miage: David Gross-Amblard, OLAP and NoSQL databases, ISTIC.
- Master: David Gross-Amblard, NoSQL databases, ISTIC.
- Master: Zoltan Miklos, Data and knowledge management (advanced course), M2 research, ISTIC
- Master: Arnaud Martin, research module on data mining and data fusion, M2 research, ENSSAT.
- Engineering school (niveau Master) Zoltan Miklos, Artificial Intelligence, ESIR (bac+4)
- Engineering school (Master level) Zoltan Miklos Project in Artificial intelligence, ESIR (bac+4)
- Engineering school (Master level) Zoltan Miklos, Data mining
- ENSAI (National School of Statistics) David Gross-Amblard, NoSQL databases
- ENS Rennes: David Gross-Amblard, "préparation à l'agrégation d'Informatique" (databases)

[2] `http://www.egc.asso.fr`

[3] `http://bdav.org`

6.2.2 Administration

- Mickaël Foursov is director of special program in Business Informatics (MIAGE)
- Yolande Le Gall is the responsable of the alternating training courses of the second year of computer sciences of the DUT
- Arnaud Martin is the director at IUT Lannion (September 2021-)
- Zoltan Miklos is the responsable of the option Information Systems, ESIR

6.2.3 Supervision

- PhD completed: Maria Massri, Gestion distribuée de grands graphes dynamiques, application à l'IOT, Zoltan Miklos. She defended her thesis in December 2022
- PhD in progress: Arthur Hoarau, Active learning of imprecise and uncertain data, Jean-Christophe Dubois, Yolande Le Gall, Arnaud Martin
- PhD in progress: Francois Mentec, Recommandation sur le recrutement de consultants et Intelligence Artificielle, Zoltan Miklos. He should defend his thesis in 2023
- PhD in progress: Zuowei Zhang, Network data mining under uncertainty using belief functions, Zhunga Liu, Arnaud Martin, cotutelle NPU, Xi'an, China, defended January 2022
- Aymen Bazouzi, Recommendations of educational resources, through graph representation learning. Supervised by Zoltan Miklos (director), Mickael Foursov and Hoel Le Capitaine (Nantes University, laboratoire LS2N), project Clara (Cominlabs)
- Erwan Vincent, Simulation and prediction of public transport services, especially, high-quality bus services, from February 2022, Zoltan Miklos (co-director), Guillaume Gravier (co-director) and Simon Malinowski (Linkmedia team). (Cifre KEOLIS Rennes), Mathieu Chambe, Zoltan Miklos (co-director), Kadi Bouatouche (co-director emeritius), Remi Cozot (Univ Litoral) Prediction of the esthetic quality of HDR images, (PhD stipend, ATER)

6.2.4 Juries

- David Gross-Amblard, PhD committee president (Univ Rennes): Johanne BAKALARA[4], "Temporal Model to explore Administrative Healthcare Databases", June 23, 2022.

6.3 Popularization

- Z. Miklos: preparation of a 2 days course on Artificial intelligence for primary and secondary degree teachers (the classes will be in March 2023), ISFEC Bretagne

- C. Thierry: "Challenge Ada Lovelace", Highschool Le Dantec, Lannion (25-26/02/2022)

[4] https://www.theses.fr/2022REN1B034

7 Bibliography

Books and Monographs

[1] A. Hameurlain, A. M. Tjoa, E. Pacitti, Z. Miklos, *Transactions on Large-Scale Data- and Knowledge-Centered Systems LI: Special Issue on Data Management - Principles, Technologies and Applications*, *Lecture Notes in Computer Science book series (LNCS)*, *13410*, Springer, 2022, https://hal.science/hal-03842737.

[2] R. Jaziri, A. Martin, M.-C. Rousset, L. Boudjeloud-Assala, F. Guillet, *Advances in Knowledge Discovery and Management, Volume 9*, *Studies in Computational Intelligence*, *1004*, Springer Internation Publishing 2022, March 2022, https://hal.archives-ouvertes.fr/hal-03614440.

Articles in referred journals and book chapters

[3] N. Andriamilanto, T. Allard, G. Le Guelvouit, A. Garel, "A Large-scale Empirical Analysis of Browser Fingerprints Properties for Web Authentication", *ACM Transactions on the Web 16*, 1, February 2022, p. 1–62, https://hal.archives-ouvertes.fr/hal-02870826.

[4] M. Chambe, R. Cozot, O. Le Meur, "Deep learning for assessing the aesthetics of professional photographs", *Computer Animation and Virtual Worlds*, July 2022, p. 14, https://hal.archives-ouvertes.fr/hal-03900933.

[5] R.-A. Cherrueau, M. Delavergne, A. van Kempen, A. Lebre, D. Pertin, J. Rojas Balderrama, A. Simonet, M. Simonin, "EnosLib: A Library for Experiment-Driven Research in Distributed Computing", *IEEE Transactions on Parallel and Distributed Systems 33*, 6, June 2022, p. 1464–1477, https://hal.inria.fr/hal-03324177.

[6] G. Lyan, D. Gross-Amblard, J.-M. Jézéquel, S. Malinowski, "Impact of Data Cleansing for Urban Bus Commercial Speed Prediction", *SN Computer Science 3*, 82, 2022, p. 1–11, https://hal.inria.fr/hal-03220449.

[7] G. Lyan, J.-M. Jézéquel, D. Gross-Amblard, R. Lefeuvre, B. Combemale, "Reasoning over Time into Models with DataTime", *Software and Systems Modeling*, December 2022, p. 1–25, https://hal.inria.fr/hal-03921928.

[8] Y. Zhang, S. Destercke, Z. Zhang, T. Bouadi, A. Martin, "On computing evidential centroid through conjunctive combination: an impossibility theorem", *IEEE Transactions on Artificial Intelligence*, June 2022, p. 1–10, https://hal.archives-ouvertes.fr/hal-03698839.

[9] Z.-W. Zhang, Z.-G. Liu, A. Martin, K. Zhou, "BSC: Belief Shift Clustering", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, p. 1–13, https://hal.archives-ouvertes.fr/hal-03816204.

[10] Z.-W. Zhang, H.-P. Tian, L.-Z. Yan, A. Martin, K. Zhou, "Learning a credal classifier with optimized and adaptive multi-estimation for missing data imputation", *IEEE Transactions on Systems, Man, and Cybernetics: Systems 52*, 7, 2022, p. 4092–4104, https://hal.archives-ouvertes.fr/hal-03271783.

Publications in Conferences and Workshops

[11] A. Hoarau, A. Martin, J.-C. Dubois, Y. Le Gall, “Imperfect Labels with Belief Functions for Active Learning”, *in: Belief Functions: Theory and Applications*, *Lecture Notes in Computer Science*, *13506*, Springer International Publishing, p. 44–53, Paris, France, October 2022, https://hal.archives-ouvertes.fr/hal-03817218.

[12] A. Martin, “Belief functions on ordered frames of discernment”, *in: 7th International Conference on Belief Functions (BELIEF2022)*, BFAS, Paris, France, October 2022, https://hal.archives-ouvertes.fr/hal-03807181.

[13] M. Massri, Z. Miklos, P. Raipin, P. Meye, “Clock-G: A temporal graph management system with space-efficient storage technique”, *in: International Conference on Data Engineering (ICDE 2022)*, Kuala Lumpur, Malaysia, May 2022, https://hal.inria.fr/hal-03621342.

[14] M. Massri, Z. Miklos, P. Raipin, P. Meye, “RTGEN : A Relative Temporal Graph GENerator”, *in: DATAPLAT workshop at the EDBT/ICDT 2022 Joint Conference*, Edinburgh, United Kingdom, March 2022, https://hal.inria.fr/hal-03609893.

[15] C. Thierry, A. Hoarau, A. Martin, J.-C. Dubois, Y. Le Gall, “Real bird dataset with imprecise and uncertain values”, *in: 7th International Conference on Belief Functions*, Paris, France, October 2022, https://hal.inria.fr/hal-03850395.

[16] H. Tian, Z. Zhang, A. Martin, Z. Liu, “Reliability-Based Imbalanced Data Classification with Dempster-Shafer Theory”, *in: 7th International Conference on Belief Functions*, PARIS, France, October 2022, https://hal.archives-ouvertes.fr/hal-03816207.