Activity Report 2016

Team LACODAM

Large Scale Collaborative Data Mining

Inria teams are typically groups of researchers working on the definition of a common project, and objectives, with the goal to arrive at the creation of a project-team. Such project-teams may include other partners (universities or research institutions).

RESEARCH CENTER
Rennes - Bretagne-Atlantique

THEME
Data and Knowledge Representation and Processing

Table of contents

Team LACODAM

Creation of the Team: 2016 January 01

Keywords:

Computer Science and Digital Science:

3. - Data and knowledge
3.1.1. - Modeling, representation
3.2. - Knowledge
3.2.2. - Knowledge extraction, cleaning
3.2.3. - Inference
3.3. - Data and knowledge analysis
3.3.1. - On-line analytical processing
3.3.2. - Data mining
3.3.3. - Big data analysis
3.4. - Machine learning and statistics
3.4.1. - Supervised learning
4.9.1. - Intrusion detection
7.1. - Parallel and distributed algorithms
8. - Artificial intelligence
8.1. - Knowledge
8.2. - Machine learning
8.6. - Decision support

Other Research Topics and Application Domains:

1.2. - Ecology
2.4. - Therapies
2.4.2. - Drug resistance
3.4.3. - Pollution
3.5. - Agronomy
4. - Energy
5.4. - Microelectronics
6.2. - Network technologies
9. - Society and Knowledge

1. Members

Research Scientists

René Quiniou [Inria, Researcher]
Torsten Schaub [University of Potsdam, Germany, International senior chair, senior Researcher, HDR]

Faculty Members

Alexandre Termier [Team leader, Univ. Rennes I, Faculty Member, Professor, HDR]
Marie-Odile Cordier [Univ. Rennes I, Faculty Member, Emeritus Professor, HDR]
Véronique Masson [Univ. Rennes I, Associate Professor]
Laurence Rozé [INSA Rennes, Associate Professor]

Thomas Guyet [Agrocampus Ouest Rennes, Associate Professor, Inria half-secondment from Sept. 2015 to Aug. 2016]
Christine Largouët [Agrocampus Ouest Rennes, Associate Professor]

Engineer
Louis Bonneau de Beaufort [Agrocampus Ouest Rennes]

PhD Students
Yann Dauxais [Univ. Rennes I]
Clément Gautrais [Univ. Rennes I]
Maël Guillemé [Energiency, granted by CIFRE, from Oct 2016]

Post-Doctoral Fellows
Benjamin Négrevergne [Inria, until Sep 2016]
Ahmed Samet [Univ. Rennes I, from May 2016]

Administrative Assistant
Marie-Noëlle Georgeault [Inria]

Others
Philippe Besnard [CNRS, Research Scientist, HDR]
Anne-Isabelle Graux [INRA, Researcher]

2. Overall Objectives

2.1. Overall Objectives

Ubiquitous data collection is providing our society with tremendous volumes of data about human, environmental and industrial activity. These ever increasing volumes of collected data hold the keys to new discoveries, both in the industrial and scientific domains. However, those keys will only be accessible to those who can make sense of such data. Making sense of data is a hard problem, requiring a good understanding of the data at hand, of the many data analysis tools and methods, and a good capacity to infer knowledge from the results of such tools. Such skills have been grouped under the umbrella term "Data Science" and lots of efforts are being done on education and research in this area. "Data Scientist" is currently the most sought after job in the US, as the demand far exceeds the number of competent professionals. Nowadays, the main problem of data science is that despite considerable improvements, it is still mostly a "manual" process: current data analysis tools still require an important human effort and know-how, making data analysis a lengthy, partial and error-prone process. This is true even for data science experts, and current approaches are mostly out of reach of non-specialists.

We claim that nowadays, Data Science is in its "Iron Age": good tools are available, however skilled craftsmen are required to use them in order to transform raw material (the data) into finished products (knowledge, decisions). We foresee that in a decade from now, we should be in an "Industrial Age" of Data Science, where more elaborate tools will alleviate a lot of the human work required in Data Science. Basic Data Science tasks will no longer require a skilled data scientist, but software tools will enable small companies or even individuals to get valuable knowledge from their data, which is not possible currently. Skilled data scientists will thus be fully available to work on the hard tasks that matter, with a drastic productivity improvement thanks to better tools doing the tedious work for them.

The objective of the Lacodam team is to considerably facilitate the process of making sense from large quantities of data, either to derive new knowledge or for making better decisions. Nowadays, this process is mostly manual, and relies on the analyst's understanding of the domain, of the data at hand and of a plethora of complex computational tools. We envision a novel generation of data analysis and decision support tools that require significantly less tedious human work, relying only on few interactions with high added value. The solutions we foresee requires to bridge data mining techniques with artificial intelligence (AI) approaches, both to take knowledge into account in a principled way, and to introduce automated reasoning techniques

in knowledge discovery workflows. Such solutions can be seen as "second order" AI tasks: they exploit AI techniques (for example, planning) in order to pilot more classical AI tasks such as data mining and decision support.

3. Research Program

3.1. Introduction

The three research axes of the Lacodam project-team are the following. First, we briefly introduce these axes, as well as their interplay:

- The first research axis is dedicated to the design of *novel pattern mining methods*. Pattern mining is one of the most important approaches to discover novel knowledge in data, and one of our strongest areas of expertise. Work in this axis will be the most fundamental of all three axes, and is expected to serve as foundations for work on the other two axes.
- The second axis tackles another aspect of knowledge discovery in data: the *interaction between the user and the system*, in order to co-discover novel knowledge. Our team has a long experience to collaborate with domain experts, and is thus especially aware of the need to improve such interaction.
- The third axis concerns *decision support*. With the help of methods from the two previous axes, our goal here is to design systems that can either help humans to take better decisions in precise applicative contexts, or to allow machines to automatically take relevant decisions in situations where extremely fast reaction time is required.

The following figure sums up the detailed work presented in the next few pages: on the sides are the three research axes of the team (X-axis) and our main applications areas (Y-axis). In the middle are colored squares that represent the precise research topics of the team that will be described in this section, placed relatively to their axis and main application area. Lines represent projects that can link several topics, and that are also connected to their main application area.

3.2. Pattern mining algorithms

Twenty years of research in pattern mining have resulted in efficient approaches to handle the algorithmic complexity of the problem. Existing algorithms are now able to efficiently extract patterns with complex structures (ex: sequences, graphs, co-variations) from large datasets. However, when dealing with large, real world datasets, these methods still output a huge set of patterns, which is impractical for human analysis. This problem is called pattern explosion. The ongoing challenge of pattern mining research is to extract fewer but more meaningful patterns. The Lacodam team is committed to solve the pattern explosion problem following four research topics:

- the design of dedicated algorithms for mining temporal patterns
- the design of flexible pattern mining approaches
- the selection of interesting data mining results
- the design of parallel pattern algorithms to ensure scalability

The originality of our contributions relies on the exploration of knowledge-based approaches whose principle is to incorporate dedicated domain knowledge (aka application background knowledge) deep into the mining process. While most of the data mining approaches are based on agnostic approaches that are designed to cope with the pattern explosion, we propose to develop data mining techniques relying on knowledge-based artificial intelligence techniques. This covers the use of structured knowledge representations, as well as reasoning methods, in combination with mining.

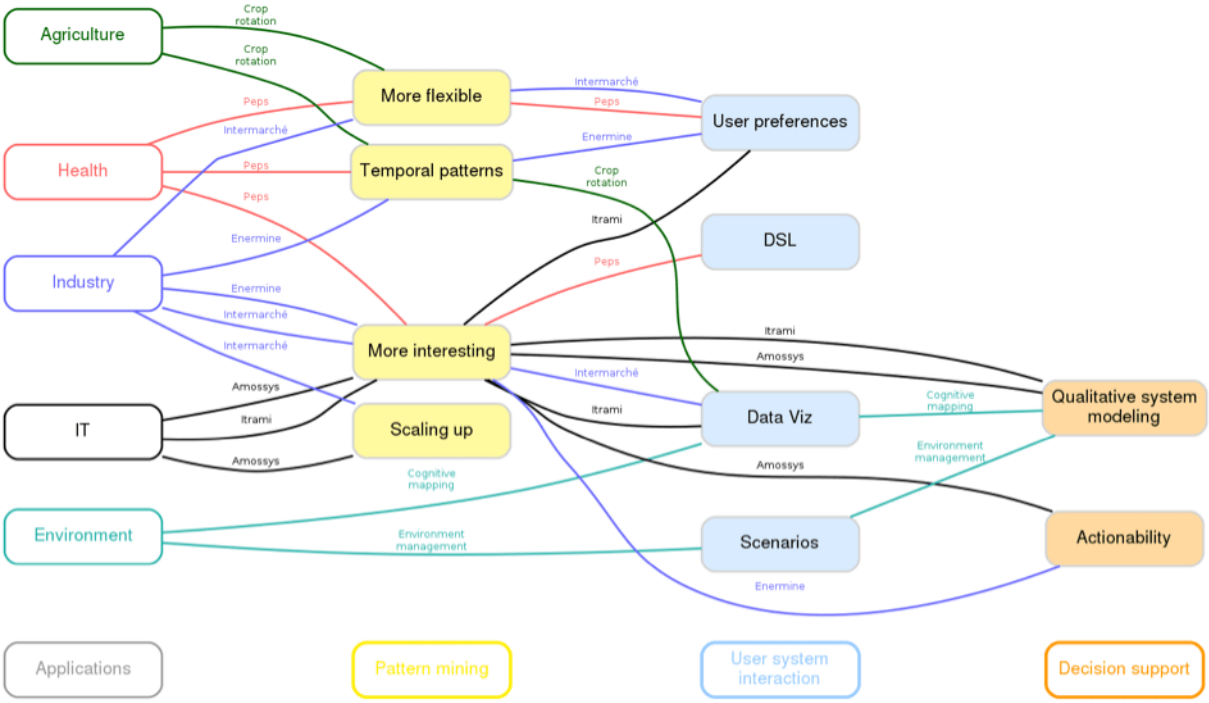

Figure 1. Lacodam research topics organized by axis and application

The first approach concerns the classical approach of pattern mining which consists in using expert knowledge to define new pattern types (and related algorithms) that can solve applicative issues. In particular, we investigate how to handle temporality in pattern representations which turns out to be important in many real world applications (in particular for decision support) and deserves particular attention.

The two other approaches aim at proposing alternative pattern mining methods to let the user incorporate, by her own, knowledge that will help define her pattern domain of interest. Flexible pattern mining approaches enable analysts to easily incorporate extra knowledge, for example domain related constraints, in order to extract only the most relevant patterns. On the other hand, the selection of interesting data mining results aims at devising strategies to filter out the results that are useless for the data analyst. Beside the challenge related to algorithmic efficiency of such approaches, we are interested in formalizing the foundations of interestingness, according to background knowledge modeled with logic knowledge representation paradigms.

Last, pattern mining algorithms are computation-intensive, it is thus important to exploit all the available computing power. Parallelism is for a foreseeable future one of the main ways to speed up computations, and we have a strong competence on the design of parallel pattern mining algorithms. We will exploit this competence in order to guarantee that our approaches scale up to the real data provided by our partners.

3.3. User/system interaction

As we pointed out before, there is a strong need to present relevant patterns to the user. This can be done by using more specific constraints, background knowledge and/or tailor-made optimization functions. Due to the difficulty of determining these elements beforehand, one of the most promising solutions is that the system and the user co-construct the definition of most relevant patterns, i.e., to have a human in the loop. This requires to have means to present intermediate results to the user, and to get user feedback in order to guide the search space exploration process in the right direction. This is an important research axis for Lacodam, which will be tackled in several complementary ways:

- Domain Specific Languages: one way to interact with the user is to propose a Domain Specific Language (DSL) tailored to the domain at hand and to the analysis tasks to perform. The challenge is to propose a DSL allowing the users to easily express the required processing workflows, to deploy those workflows for mining large volumes of data and to offer as much automation as possible.
- What if / What for scenarios: we are also investigating the use of scenarios to query results from data mining processes, as well as other complex processes such as complex system simulations or model predictions. Such scenarios are answers to questions of the type “what if [situation]” or “what [should be done] for [expected outcome]”.
- User preferences: in exploratory analysis, users often do not have a precise enough idea of what they want, and are not able to formulate such queries. Lacodam is thus investigating simple ways for letting users express their interests and preferences, either during the mining process to guide the search space exploration, or after, to help in getting the most relevant results.
- Data visualization: most of the research directions presented in this document require users to examine patterns at some point. The output of most pattern mining algorithms is simply a (long) list of patterns. While this presentation can be sufficient in some applications, it is often not enough to provide a complete understanding, especially for non-experts in pattern mining. A transversal research topic that we want to develop in Lacodam is to propose data visualization techniques adequate to understanding output results. Numerous (failed) experiments have shown that data mining and data visualization are fields which require distinct skills, where researchers in one field usually do not make significant advances in the other field (this is detailed in [Keim 2010]). Thus, our strategy is to establish collaborations with prominent data visualization teams for this line of research, with a long term goal to recruit a specialist in data visualization if the opportunity arises.

3.4. Decision support

Patterns, especially predictive sequential patterns, resulting from mining a dataset have often a direct application in diagnosis. Lacodam inherits from the former Dream team a strong background in decision support systems, with an internationally recognized expertise in diagnosis. This AI subfield is concerned with determining if a system is operating normally or not, and if the system is in an abnormal state, to determine the cause of the faulty behavior. The considered system can as well be an agro- or eco-system, a software system or an animal or human being, as well.

The increasing volumes of data coming from a wide range of different systems (ex: sensor data from agro-environmental systems, log data from software systems, biological data coming from health monitoring systems) show that it is possible to gather more and more observations for such systems. Thus, it should be possible to exploit such observations to help human or software agents to take better decisions. Hence, while keeping the strong interest on decision support (and especially diagnosis) that existed in Dream, Lacodam adds the idea that the decision support systems should take advantage of the huge volumes of data available. This third and last research axis is thus a meeting point for all members of the team, as it requires to integrate AI techniques of traditional decision support systems with results from data mining techniques.

Two main research axes are investigated in Lacodam:

- Diagnosis-based approaches. We are exploring how to integrate knowledge found from pattern mining approaches, possibly with the help of interactive methods, into the qualitative models. The goal of such work is to partly automate the construction of the model, which can require a lot of human effort otherwise.
- Actionable patterns and rules. In many settings of "exploratory data mining", the actual interest of a pattern is hard to assess, as it may be hard to measure or may be subjective (resulting from introducing the user in the mining process). However, there exist applications where once patterns are found, there are well defined measures to define what this pattern will bring to the user. Further, patterns and rules that can lead to actual actions beneficial to the user are called actionable patterns. Such actionable patterns and rules are especially important for industry.

3.5. Long-term goals

The following perspectives are at the convergence of the three research axes presented before, and can be seen as the ideal towards which our efforts tend:

- Automating data science workflow discovery. The current methods for extracting knowledge from data and building decision support systems require a lot of human effort. Our three research axes aim at alleviating this effort, by devising methods that are more generic and by improving the interaction between the user and the system. An ideal solution would be that the user could forget completely about the existence of pattern mining or decision support methods. Instead the user would only loosely specify her problem, while the system would construct for her various data science / decision support workflows, possibly further refined via interactions.

  We consider that this is a second order AI task, where AI techniques such as planning are used to explore the workflow search space, the workflow itself being composed of data mining and/or decision support components. This is a strategic evolution for data science endeavors, were the demand far exceeds the available human skilled manpower.
- Logic argumentation based on epistemic interest. Having increasingly automated approaches will require better and better ways to handle the interactions with the user. Our second long term goal is to explore the use of logic argumentation as an interaction tool between users and a data analysis tool. Alongside visualization and interactive data mining tools, it can be a way for users to query in an intuitive manner both the results and the way they were obtained. Such querying can also help the expert to reformulate her query in an interactive analysis setting.

This research direction continues the work on "epistemic interest" presented before. Its goal is to exploit principles of interactive data analysis in the context of epistemic interest measures. Logic argumentation [Besnard 2014] can be a natural tool for interactions between the user and the system: display of possibly exhaustive list of arguments, relationships – whether reinforcement, compatibility or conflict – between arguments, variable degrees of arguments, and possible solutions for argument conflicts.

The first step is to define a formal argumentation framework for explaining data mining results. This implies to continue theoretical work on the foundations of argumentation in order to identify the most adapted framework (either existing or a new one to be defined). Logic argumentation may be implemented and deeply explored in ASP, allowing us to build on our expertise in this logic language.

- Collaborative feedback and knowledge management. We are convinced that improving the data science process, and possibly automating it, will rely at some point in the near future on the vast feedback that can be obtained by communities of user seamlessly collaborating over the web. Consider for example what has been achieved by collaborative platforms such as StackOverflow: it has become the reference site for any programming question.

  Data science is a more complex problem than programming, as in order to get help from the community, the user has to share her data and workflow, or at least some parts of them. This raises obvious privacy issues that may prevent this idea to succeed. As our research on automating the production of data science workflows should enable more people to have access to data science results, we are interested to investigate the design of collaborative platforms to exchange expert advices over data, workflows and analysis results, with an aim at exploiting this human feedback to improve the automated system with machine learning.

4. Application Domains

4.1. Introduction

The current period is extremely favorable for teams working in Data Science and Artificial Intelligence, and Lacodam is no exception. We are eager to see our work applied in real world applications, and have thus an important activity in maintaining strong ties with industrials partners concerned with marketing and energy as well as public partners working in health, agriculture and environment.

4.2. Industry

We present below our industrial collaborations. Some are well established partnerships, while others are more recent collaborations with local industries that wish to reinforce their Data Science R&D with us (e.g. STMicroelectronics, Energiency, Amossys).

- **Execution trace analysis for SOC debugging (STMicroelectronics)**. We have an ongoing collaborations with STMicroelectronics, which is one of the world top-5 electronic chip makers. Nowadays, set-top boxes, smartphones or onboard car computers are powered by highly integrated chips called System-on-Chip (SoC). Such chips contain on a single die processing units, memories, IO units and specialized accelerators (such as audio and video encoding/decoding). Programming SoC is a hard task due to their inherent parallelism, leading to subtle bugs when several components do not deliver their results within a given time frame. Existing debuggers and profilers are ill-adapted in this case because of their high intrusivity that modifies the timings. Hence the most used technique is to capture a trace of the execution and analyze it post-mortem. While Alexandre Termier was in Grenoble he initiated several works for analyzing such traces with pattern mining techniques, which he is now pursuing with his colleagues of the Lacodam project-team.

- **Resource consumption analysis for optimizing energy consumption and practices in industrial factories (Energiency)**. In order to increase their benefits, companies introduce more and more sensors in their factories. Thus, the resource (electricity, water, etc.) consumption of engines, workshops or factories are recorded in the form of times series or temporal sequences. The person who is in charge of resource consumption optimization needs better software than classical spreadsheets. He/she needs effective decision-aiding tools with statistical and artificial intelligence knowledge. The start-up Energiency aims at designing and offering such pieces of software for analyzing energy consumption. The starting CIFRE PhD thesis of Maël Guillemé aims at proposing new approaches and solutions from the data mining field to tackle this issue.
- **Security (Amossys)**. Current networks are faced with an increasing variety of attacks, from the classic « DDoS » that makes a server unusuable for a few fours, to advanced attacks that silently infiltrate a network and exfiltrate sensitive information monthes or even years later. Such intrusions, called APT (Advanced Persistent Threat) are extremely hard to detect, and this will become even harder as most communications will be encrypted. A promising solution is to work on "behavioral analysis", by discovering patterns based on the metadata of IP-packets. Such patterns can relate to an unusual sequencing of events, or to an unusual communication graph. Finding such complex patterns over a large volume of streaming data requires to revisit existing stream mining algorithms to dramatically improve their throughput, while guaranteeing a manageable false positive rate. We are collaborating on this topic with the Amossys company and the Emsec team of Irisa through the co-supervision of a CIFRE PhD (located in the Emsec team). Our goal is to design novel anomaly detection methods that can detect APT, and that scales on real traffic volumes.
- **Market basket data analysis (Intermarché) and multi-channel interaction data analysis (EDF) for better Customer Relationship Management (CRM)**. An important application domain of data mining for companies that deal with large numbers of customers is to analyze customer interaction data, either for marketing purposes or to improve the quality of service. We have activities in both settings. In the first case, we collaborate with a major french retailer, Intermarché, in order to detect customer churn by analyzing market basket data. In the second case, we collaborate with the major french power supplier, EDF, to discover actionable patterns for CRM aiming at avoiding reaching undesirable situations from logs of user interactions with the company (web clicks, phone calls, etc.).

4.3. Health

- **Care pathways analysis for supporting pharmaco-epidemiological studies**. Pharmaco-epidemiology applies the methodologies developed in general epidemiology to answer to questions about the uses and effects of health products, drugs [20], [19] or medical devices [17], on population. In classical pharmaco-epidemiology studies, people who share common characteristics are recruited to build a dedicated prospective cohort. Then, meaningful data (drug exposures, diseases, etc.) are collected from the cohort within a defined period of time. Finally, a statistical analysis highlights the links (or the lack of links) between drug exposures and outcomes (*e.g.* adverse effects). The main drawback of prospective cohort studies is the time required to collect the data and to integrate it. Indeed, in some cases of health product safety, health authorities have to answer quickly to pharmaco-epidemiology questions.

  New approaches of pharmaco-epidemiology consist in using large EHR (Electronic Health Records) databases to investigate the effects and uses (or misuses) of drugs in real conditions. The objective is to benefit from nationwide available data to answer accurately and in a short time pharmaco-epidemiological queries for national public health institutions. Despite the potential availability of the data, their size and complexity make their analysis long and tremendous. The challenge we tackle is the conception of a generic digital toolbox to support the efficient design of a broad range of pharmaco-epidemiology studies from EHR databases.

  We propose to use pattern mining algorithm and reasoning techniques to analyse the typical care pathways of specific groups of patients.

To be able to answer the broad range of pharmaco-epidemiological queries from national public health institutions, the PEPS [1] platform exploits, in secondary use, the French health cross-schemes insurance system, called SNIIRAM. The SNIIRAM covers most of the French population with a sliding period of 3 past years. The main characteristics of this data warehouse are described in [18]. Contrary to local hospital EHR or even with other national initiatives, the SNIIRAM data warehouse covers a huge population. It makes possible studies on unfrequent drugs or diseases in real conditions of use. To tackle the volume and the diversity of the SNIIRAM data warehouse, a research program has been established to design an innovative toolbox. This research program is focused first on the modeling of care pathways from the SNIIRAM database and, second, on the design of tools supporting meaningful insights extraction about massive and complex care pathways by clinicians. In such database a care pathway is an individual sequence of drugs exposures, medical procedures and hospitalizations.

4.4. Agriculture and environment

- **Dairy farming**. The use and analysis of data acquired in dairy farming is a challenge both for data science and for animal science. Its goal is to improve farming conditions (health, welfare and environment) as well as farmers' income. Nowadays, animals are monitored by multiple sensors giving a wealth of heterogeneous data (ex: temperature, weight, milk composition...). Current techniques used by animal scientists focus mostly on mono-sensor approaches. The dynamic combination of several sensors could provide new services and information useful for dairy farming. A PhD thesis will begin soon to study such combinations of sensors and to investigate data mining methods, especially pattern mining algorithms. The challenge is to design new algorithms taking into account the data heterogeneity, coming both from their nature and the different time scales involved, and to produce patterns that are actually useful for dairy farming. This thesis will be an original and important contribution to the new challenge of the IoT (Internet of Things) and will interest domain actors to find new added value to a global data analysis. The PhD thesis will take place in an interdisciplinary setting bringing together computer scientists from Inria and animal scientists from INRA, both located in Rennes.

  Similar problems are investigated with the veterinary department of the University of Calgary in the context of cattle monitoring from multiple sensors placed on calves for the early detection of diseases.
- **Optimizing the nutrition of individual sow**. Another direction for further research is to combine data flow with prediction models in order to learn nutrition strategies. We are currently starting a project with INRAon the nutritional requirements and the optimal diet to be supplied to individual lactating sow. The research issue will be to develop decision algorithms for the determination of the optimal ration (amount and composition) to be fed to a given sow, on a given day, considering all the information available (real-time observation data flow and historical data). Issues concern the design of an incremental learning algorithm that will compute the animal profile and how to determine the best feeding plan. Efficiency issues of developed algorithms will also be considered since the proposed software should work in real-time on the automated feeder.
- **Ecosystem modeling and management**. Ongoing research on ecosystem management includes modelling of ecosystems and anthroprogenic pressures, with a special concern on the representation of socio-economical factors that impact human decisions. A main research issue is how to to represent these factors and how to integrate their impact on the ecosystem simulation model. This work is an ongoing cooperation with ecologists from the Marine Spatial Ecology of Queensland University, Australia and from Agrocampus Ouest.

5. Highlights of the Year

[1] PEPS: Pharmaco-Epidémiologie et Produits de Santé – Pharmacoepidemiology of health products

5.1. Highlights of the Year

- This year, we are extremely proud to have a total of 4 papers accepted at the IJCAI conference, the rank A+ conference on Artificial Intelligence.
- Another highlight of this year is that following the end of the former team, namely Dream, we could propose in 2016 a new team project, namely Lacodam, and follow smoothly all steps of the Inria project-team creation protocol. While the team is not officially created as of December 2016, our project has been positively evaluated both by Inria members and by international experts, and is thus likely to be created in early 2017.

6. New Software and Platforms

6.1. EcoMata

FUNCTIONAL DESCRIPTION

The EcoMata tool-box provides means for qualitative modeling and ecosystem exploration and for assisting the design of environmental guidelines. We have proposed a new qualitative approach for ecosystem modeling based on timed automata (TA) formalism combined to a high-level query language for exploring scenarios.

- Participants: Marie-Odile Cordier, Yulong Zhao, Christine Largouët and Thomas Guyet
- Contact: Christine Largouët

6.2. PaturMata

KEYWORDS: Bioinformatics - Biology
SCIENTIFIC DESCRIPTION

The Paturmata tool-box provides means for qualitative modeling and exploring agrosystems, specifically management of herd based on pasture. The system is modelled using a hierarchical hybrid model described in timed automata formalism.
FUNCTIONAL DESCRIPTION In the PaturMata software, users can create a pasture system description by entering herds and plots information. For each herd, the only parameter is the number of animals. For each plot, users should enter the surface, the density, the herb height, the distance to the milking shed, a herb growth profile and an accessibility degree.

Users then specify pasturing and fertilization strategies. Finally, users can launch a pasture execution. PaturMata displays the results and a detailed trace of pasture. Users can launch a batch of different strategies and compare the results in order to find the best pasture strategy.

PaturMata is developed in Java (Swing for the GUI) and the model-checker that is called for the timed properties verification is UPPAAL .

- Participants: Christine Largouët, Marie-Odile Cordier, Yulong Zhao
- Contact: Christine Largouët

6.3. QTempIntMiner

KEYWORDS: Data mining - Health - Medical - Physiology - Temporal information
SCIENTIFIC DESCRIPTION

QTempIntMiner: the QTempIntMiner (Quantitative Temporal Interval Miner) data mining software implements several algorithms presented in [46] and [3] (QTIAPriori and QTIPrefixSpan). The software is mainly implemented in Matlab. It uses the Mixmod toolbox [33] to compute multi-dimensional Gaussian distributions. The main features of QTempIntMiner are:

- a tool for generating synthetic noisy sequences of temporal events,
- an implementation of the QTempIntMiner , QTIAPriori and QTIPrefixSpan algorithms,
- a graphical interface that enables the user to generate or import data set and to define the parameters of the algorithm and that displays the extracted temporal patterns.
- a sequence transformer to process long sequences of temporal events. Long sequences are transformed into a database of short temporal sequences that are used as input instances for the available algorithms.

The software includes one new algorithm based on the separation of the set of intervals to extract more efficiently but less accurately the time interval in temporal patterns. This new algorithm version is still under evaluation on simulated and real datasets (care pathways).

- Participants: Thomas Guyet and René Quiniou
- Partner: AGROCAMPUS
- Contact: Thomas Guyet
- URL: http://people.irisa.fr/Thomas.Guyet/QTempIntMiner/

7. New Results

7.1. Introduction

In this section, we organize our contributions over three main research topics:

- Mining different kinds of patterns, from 7.2 to 7.9
- Data mining and decision support with ASP, from 7.10 to 7.13
- Model-based diagnosis, from 7.13 to 7.15.

7.2. Customer Purchase Signatures: a New Model in Grocery Retail Context

Participants: Clément Gautrais, Peggy Cellier [Lis], Thomas Guyet, René Quiniou, Alexandre Termier.

In the retail context, there is an increasing need for understanding individual customer behavior in order to personalize marketing actions. We propose the novel concept of customer signature, that identifies a set of important products that the customer refills regularly. Both the set of products and the refilling time periods give new insights on the customer behavior. Our approach is inspired by methods from the domain of sequence segmentation, thus benefiting from efficient exact and approximate algorithms. Experiments on a real massive retail dataset show the interest of the signatures for understanding individual customers (under submission to PAKDD 2017 conference).

This new model is used to detect and explain customer defection in a grocery retail context from the evolution of each customer basket content. It therefore provides actionable knowledge for the retailer at an individual scale. In addition, this model is able to identify customers that are likely to defect in the future months [16].

7.3. Discriminant Chronicles for Care Pathway Analysis

Participants: Yann Dauxais, Thomas Guyet, David Gross-Amblard [Druid], André Happe [Brest University Hospital].

A care pathway is a sequence of events (drugs deliveries, hospitalisation, etc) extracted from medical databases (see section 4.3 for details). In some studies, each patient is labeled by a class (*e.g.* died or not died). This information can be taken into account for the discriminant analysis of care pathways. This year, our objective was to extract discriminant patterns from a dataset of care pathways that can discriminate patients on their labels. To this end we introduced the new task of discriminant chronicle mining. Conceptually, a chronicle is a graph whose vertices are events and edges represent quantitative time constraints between events. We also proposed DCM, an algorithm dedicated to discriminant chronicles mining. This algorithm is based on rule learning methods to extract the temporal constraints. Computational performances and discriminant power of extracted chronicles are evaluated on artificial and real data.

The paper describing this work has been accepted in the french national conference on data mining (EGC 2017) [4] and is nominated for the best paper award.

7.4. Identifying Genetic Variant Combinations using Skypatterns

Participants: Alexandre Termier, Hoang-Son Pham [Genscale], Dominique Lavenier [Genscale].

Identifying variant combination association with disease is a bioinformatics challenge. This problem can be solved by discriminative pattern mining that uses a statistical function to evaluate the significance of individual biological patterns. There is a wide range of such measures. However, selecting an appropriate measure as well as a suitable threshold in some specific practical situations is a difficult task. In this work, we propose to use the skypattern technique which enables using combinations of measures to evaluate the importance of variant combinations without having to select a given measure and a fixed threshold (Pareto frontier). Experiments on several real variant datasets demonstrates that the skypattern method effectively identifies the risk variant combinations related to diseases [13].

7.5. Steady Patterns

Participants: Alexandre Termier, Willy Ugarte [UGA Grenoble], Miguel Santana [STMicroelectronics].

Skypatterns are an elegant answer to the pattern explosion issue, when a set of measures can be provided. Skypatterns for all possible measure combinations can be explored thanks to recent work on the skypattern cube. However, this leads to too many skypatterns, where it is difficult to quickly identify which ones are more important. First, we introduce a new notion of pattern steadiness [14] which measures the conservation of the skypattern property across the skypattern cube, allowing to see which are the "most universal" skypatterns. Then, we extended this notion to partitions of the dataset, and show in our experiments that this both allows to discover especially stable skypatterns, and identify interesting differences between the partitions.

7.6. Dense Bag-of-Temporal-SIFT-Words for Time Series Classification

Participants: Adeline Bailly [IRISA/Obelix], Laetitia Chapel [IRISA/Obelix], Thomas Guyet, Simon Malinowski [LinkMedia], Romain Tavenard [IRISA/Obelix].

The SIFT framework has shown to be effective in the image classification context. In [15], we designed a Bag-of-Words approach based on an adaptation of this framework to time series classification. It relies on two steps: SIFT-based features are first extracted and quantized into words; histograms of occurrences of each word are then fed into a classifier. In this work, we investigate techniques to improve the performance of Bag-of-Temporal-SIFT-Words: dense extraction of keypoints and different normalizations of Bag-of-Words histograms. Extensive experiments show that our method significantly outperforms nearly all tested standalone baseline classifiers on UCR datasets.

7.7. Comparing Symbolic and Statistical Classifiers on Energy Consumption Data

Participant: Benjamin Négrevergne.

During his Inria Carnot postdoc, Benjamin Négrevergne aimed at testing various data mining and machine learning methods on energy consumption data from the Energiency startup. Two symbolic methods developed in Lacodam were evaluated: QTempIntMiner and discriminant chronicle mining. While QTempIntMiner was shown to be ill-adapted in this setting, discriminant chronicle mining gave promising results. These results were evaluated in collaboration with our industrial partner. We also shown the interest of other methods: Hidden Markov Models and Gaussian processes. An internal report has been written to relate the results.

7.8. Detecting Strategic Moves in HearthStone Matches

Participants: Boris Doux [M1 intern], Clément Gautrais, Benjamin Negrevergne.

In this work, we demonstrate how to extract strategic knowledge from gaming data collected among players of the popular video game HearthStone. Our methodology is as follows. First we train a series of classifiers to predict the outcome of the game during a match, then we demonstrate how to spot key strategic events by tracking sudden changes in the classifier prediction. This methodology is applied to a large collection of HeathStone matches that we have collected from top ranked European players. Expert analysis shows that the events identified with this approach are both important and easy to interpret with the corresponding data [12].

7.9. Towards Visualizing Hidden Structures

Participants: Rémy Dautriche [STMicroelectronics], Alexandre Termier, Renaud Blanch [UGA Grenoble], Miguel Santana [STMicroelectronics].

There is an increasing need to quickly understand the contents of log data. A wide range of patterns can be computed and provide valuable information: for example existence of repeated sequences of events or periodic behaviors. However pattern mining techniques often produce many patterns that have to be examined one by one, which is time consuming for experts. On the other hand, visualization techniques are easier to understand, but cannot provide the in-depth understanding provided by pattern mining approaches. Our contribution is to propose a novel visual analytics method that allows to immediately visualize hidden structures such as repeated sets/sequences and periodicity, allowing to quickly gain a deep understanding of the log [3].

7.10. Knowledge-based Sequence Mining with ASP

Participants: Thomas Guyet, René Quiniou, Torsten Schaub.

We have introduced a framework for knowledge-based sequence mining, based on Answer Set Programming (ASP) [10], [5]. We begin by modeling the basic task and refine it in the sequel in several ways. First, we show how easily condensed patterns can be extracted by modular extensions of the basic approach. Second, we illustrate how ASP's preference handling capacities can be exploited for mining patterns of interest. In doing so, we demonstrate the ease of incorporating knowledge into the ASP-based mining process. To assess the trade-off in effectiveness, we provide an empirical study comparing our approach with a related sequence mining mechanism.

7.11. Packing Graphs with ASP for Landscape Simulation

Participants: Thomas Guyet, Yves Moinard, Jacques Nicolas [Dyliss], René Quiniou.

This work [6] describes an application of Answer Set Programming (ASP) to crop allocation for generating realistic landscapes. The task is to optimally cover a bare landscape, represented by its plot graph, with spatial patterns describing local arrangements of crops. This problem belongs to the hard class of graph packing problems and is modeled in the framework of ASP. The approach provides a compact and elegant solution to the basic problem and at the same time allows extensions such as a flexible integration of expert knowledge. Particular attention is paid to the treatment of symmetries, especially due to sub-graph isomorphism issues. Experiments were conducted on a database of simulated and real landscapes. Currently, the approach can process graphs of medium size, a size that enables studies on real agricultural practices.

7.12. Care Pathway Analysis with ASP Sequence Mining

Participants: Ahmed Samet, Benjamin Négrevergne, Thomas Guyet.

This line of work aims at applying our ASP encoding for sequential pattern mining to care pathway analysis (see section 4.3 for applicative objectives). This year, we proposed an approach of meaningful rare sequential pattern mining based on the declarative programming paradigm of Answer Set Programming (ASP). The setting of rare sequential pattern mining is introduced. To cope with the huge amount of meaningless rare patterns, our ASP approach provides an easy manner to encode expert constraints on expected patterns. Encodings are presented and quantitatively compared to a procedural baseline. An application on care pathways analysis illustrates the qualitative interest of expert constraints encoding.

This work has been submitted to the PAKDD 2017 conference.

7.13. ASP and Diagnosis

Participants: Christine Largouët, Laurence Rozé.

A new approach for performing diagnosis with ASP has been explored. The system is described by automata and implemented in an ASP program whose task is to find trajectories compatible with observations. The experimentation is carried out on benchmarks already used for the diagnosis problem using SAT. These benchmarks consider different levels of difficulty and number of faults (from one to twenty) and three types of observations: timestamped observations, totally ordered observations and partially ordered observations. The results were good both for dated and for totally ordered sequences of observations, whereas work needs to be still improved for the partial ordered observation case.

7.14. Searching for Cost-Optimized Strategies. Application to Temporal Planning and Agricultural System

Participants: Christine Largouët, Marie-Odile Cordier.

We consider a system modeled as a set of interacting components evolving along time according to explicit timing constraints. The decision making problem consists in selecting and organizing actions in order to reach a goal state in a limited time and in an optimal manner, assuming actions have a cost. We propose to reformulate the planning problem in terms of model-checking and controller synthesis such that the state to reach is expressed using a temporal logic. We have chosen to represent each agent using the formalism of Priced Timed Game Automata (PTGA) and a set of knowledge. PTGA is an extension of Timed Automata that allows the representation of cost on actions and the definition of a goal (to reach or to avoid). A first paper describes two algorithms designed to address the planning problem on a network of agents and proposes a practical implementation using model-checking tools that shows promising results on an agricultural application: a grassland based dairy production system [9]. Another paper describes the expressivity of this approach on the classical Transport Domain which is extended in order to include timing constraints, cost values and uncontrollable actions. This work has been implemented and performances evaluated on benchmarks [8].

7.15. Integrating Socio-Economic Drivers in an Explicit-Time, Qualitative Fisheries Model

Participant: Christine Largouët.

EcoMata is an explicit-time, qualitative modelling tool for assessing the ecosystem impacts of fishing and evaluating options for fishery management. The model is being developed further by integrating simple socio-economic drivers in the fishery system. Specifically, we have introduced a new module of automata that describes the profits associated to a specific fishing intensity and specific timing. This new module allows the evaluation of management strategies that are economically viable. The approach is illustrated on a coral-reef fishery in the Pacific that has been the focus of previous modelling work. [7].

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. SocTrace: analysis of SOC traces

Participant: Alexandre Termier.

SoCTrace is a FUI project led by STMicroelectronics, with the companies ProbaYes and Magilem, University Grenoble Alps and Inria Rhône-Alpes. Its goal is to provide an integrated environment for storing and analyzing execution traces. In this project, we are working on data mining techniques for analyzing the traces, and on the use of ontologies to enable querying traces with a higher level of abstraction.

8.1.2. ITRAMI: Interactive Trace Mining

Participant: Alexandre Termier.

ITRAMI is a Nano2017 project. Such projects are designed to support joint research efforts between STMicrolectronics and academic partners in the domain of embedded systems. Alexandre Termier is the PI of this project whose goal is to design novel data mining methods for interactive analysis of execution traces. Such methods aim at considerably reducing the time that STMicroelectronics developers spend at understanding, debugging and profiling applications running on STMicrolectronics chips. The project work is done at University Grenoble Alps, in collaboration with Lacodam researchers. Two contractual staff are working on the project in Grenoble: Willy Ugarte as a postdoc, and Soumaya Ben Alouane as an engineer.

8.2. Bilateral Grants with Industry

Maël Guillemé has obtained a CIFRE PhD grant with the Energiency startup, supervised by V. Masson and L. Rozé. The goal of Maël Guillemé's thesis is to propose new approaches for improving industrial energy performance and aims at integrating both numerical and symbolic attributes. A master 2 internship explored in 2016 a first approach based on an algorithm proposed by Shokoohi and al, but with several improvements: avoid data normalisation, detect patterns as fast as possible, enhance functions like distance and score.

Another CIFRE PhD has started, this time with the Amossys company, specialized in cyber-security. This is the PhD of Alban Siffer, located in the EMSec team of IRISA and co-supervised between EMSec (P.A. Fouque) and Lacodam (A. Termier, C. Largouët). The goal of this PhD is to propose new methods for intrusion detection in networks. The originality is to only consider IP flow as input (metadata of packets and not packet contents), requiring to detect intrusion via unusual traffic patterns.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. SePaDec: Declarative approaches for Sequential Pattern mining

Participants: Benjamin Negrevergne, Thomas Guyet, Ahmed Samet, Alexandre Termier.

The SEPADEC project is funded by the Region Bretagne. It aims at exploring the application of declarative pattern mining (more especially ASP) in the field of care pathway analysis. The first objective was to model knowledge from the data to enrich the raw data with medical expert knowledge and to develop a toolbox that smoothly integrates both expert knowledge and declarative pattern mining.

9.2. National Initiatives

9.2.1. ANR

9.2.1.1. #DigitAg: Digital agriculture

Participants: Alexandre Termier, Véronique Masson, Christine Largouët, Anne-Isabelle Graux.

#DigitAg is a “Convergence Institute” dedicated to the increasing importance of digital techniques in agriculture. Its goal is twofold: first, make innovative research on the use of digital techniques in agriculture in order to improve competitiveness, preserve environment, and offer correct living conditions to farmers. Second, through education prepare future farmers and agricultural policy makers to successfully exploit such technology.

While #DigitAg is based on Montpellier, Rennes is a satellite of the institute focused on cattle farming. Lacodam is involved in the “data mining” challenge of the institute, that A. Termier co-leads. He is also the representative of Inria in the steering comittee of the institute.

The interest for the team is to design novel methods to analyze and represent agricultural data, which are challenging because they are both heterogeneous and multi-scale (both spatial and temporal).

9.2.2. National Platforms

9.2.2.1. PEPS: Pharmaco-epidemiology for Health Products

Participants: Yann Dauxais, Thomas Guyet, Véronique Masson, René Quiniou, Alexandre Termier.

The PEPS project (Pharmaco-epidemiology des Produits de Santé) is funded by ANSM (national agency for health security). The project leader is E. Oger from the clinical investigation center CIC-1414 INSERM/CHU Rennes. The other partners located in Rennes are the Institute of Research and Technology (IRT) B<>Com, EHESP and the LTSI. The project started in january 2015 and is funded for 4 years.

The PEPS project has two parts: the clinical studies and a research program dedicated to the development of innovative tools for pharmaco-epidemiological studies with medico-administrative databases.

Our contribution to this project will be to propose pattern mining algorithm and reasoning techniques to analyse the typical care pathways of specific groups of insured patients.

9.3. International Initiatives

9.3.1. Inria International Partners

9.3.1.1. Informal International Partners

9.3.1.1.1. University of Calgary: Monitoring cattle in big herds with multiple sensors

Participant: René Quiniou.

The state of Alberta produces a significant part of the beef meat in Canada. Big farms feeds up around 40.000 bull calves in feedlots grouping 200-300 animals. Diseases such as Bovine Respiratory Diseases (BRD) are frequent and may propagate quickly in such conditions. So, it is important to detect as soon as possible when an animal is sick. We are collaborating with the Department of Production Animal Health, University of Calgary for designing monitoring systems able to generate early alarms when an animal is sick. Precisely, we are studying the properties of new sensors and their aptitude to provide relevant data for BRD detectors.

9.3.1.1.2. University of Potsdam: preferences in mining with ASP

Participant: Thomas Guyet.

The research group “knowledge processing and information systems" of the University of Potsdam, so called Potascco group, develops a collection of tools and programs for Answer Set Programming such as the clingo solver or the ASPRIN system, developed by J. Romero to handle preferences on ASP models. They have strong expertise in problem encoding with ASP. In addition to T. Schaub Inria position, we initiate some collaborations with other members of the Potascco group in order to strengthen our relationships.

9.4. International Research Visitors

9.4.1. Research Stays Abroad

Thomas Guyet spent a month (may 2016) in the team leaded by Prof. Torsten Schaub in the university of Potsdam.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

- Local chair of the Mini-symposium on instant data mining, interactive data mining, preference-based pattern mining, Rennes, October 26-28, 2016 (A. Termier)
- Organization chair (T. Guyet) and program committee members (T. Guyet, R. Quiniou) of GAST workshop at EGC 2017 and at EGC 2016.

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

EGC 2017 (T. Guyet, R. Quiniou, A. Termier).

ICDM 2016 (A. Termier)

KDD 2016 (A. Termier)

IJCAI 2016 (T. Guyet, B. Negrevergne)

ECAI 2016 (T. Guyet)

KR 2016 (T. Guyet)

RFIA/CNIA 2016 (T. Guyet)

APIA 2016 (C. Largouët)

10.1.2.2. Reviewer

R. Quiniou: KDD 2016, ECAI 2016, IJCAI 2016, ICDM 2016, CNIA 2016, AKDM7

T. Guyet: KDD 2016, ECAI 2016, IJCAI 2016, ICDM 2016, RFIA/CNIA 2016, KR 2016, AALTD16, SimBig 2016

A. Termier: IJCAI 2016

Y. Dauxais, C. Gautrais: IJCAI 2016, ICDM 2016, KDD 2016, EGC 2016

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

T. Guyet: RIA (Revue d'Intelligence Artificielle)

10.1.3.2. Reviewer - Reviewing Activities

A. Termier: Data Mining and Knowledge Discovery, Knowledge and Information Systems, Discrete Applied Mathematics

T. Guyet: Journal of Biomedical Informatics, AI Communications, ACM Computing Surveys, TKDE

C. Gautrais: Data Mining and Knowledge Discovery

10.1.4. Invited Talks

A. Termier gave an invited talk at the SMiLe 2016 workshop.

C. Largouët gave a invited talk at the UMR-AMURE Seminar on the 15th september 2016,

10.1.5. Leadership within the Scientific Community

A. Termier is the representative of Inria for the #DigitAg Convergence Institue

T. Guyet is member of the AFIA board (since october 2011). AFIA is the french chapter of ECCAI, the European Association for Artificial Intelligence.

10.1.6. Scientific Expertise

Evaluation of a project proposal for the ANR: R. Quiniou

Evaluation of a project proposal for the Medicen competitive cluster: T. Guyet

10.1.7. Research Administration

Member of INRA CEI (Engineers Evaluation Committee): T. Guyet

Member of the scientific board of department EA of INRA: A. Termier

Member of the scientific board of Agrocampus Ouest - COREGE: C. Largouët.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Many members of the project-team Lacodam are also faculty members and are actively involved in computer science teaching programs in ISTIC, INSA and Agrocampus-Ouest. Besides these usual teachings Lacodam is involved in the following programs:

Master 2 Module DSS: Apprentissage sur des données séquentielles symboliques, 10 h, M2, Istic University of Rennes 1 (R. Quiniou)

Master 2 Géoinformation, Agrocampus Ouest Rennes (L. Bonneau, T. Guyet)

Master 2 Artificial Intelligence, Agrocampus Ouest Rennes (Louis Bonneau, C. Largouët)

Master 1 Scientific Programming, Data Management, Python Programming, Agrocampus Ouest Rennes (C. Largouët)

10.2.2. Supervision

PhD in progress: Maël Guillemé, “New data mining approaches for improving energy consumption in factory"”, october 3rd 2016, Alexandre Termier, Véronique, Masson and Laurence Rozé

PhD in progress: Clément Gautrais, “Mining massive data from client purchases”, october 1st 2015, Alexandre Termier, Peggy Cellier, Thomas Guyet and René Quiniou

PhD in progress: Yann Dauxais, "“Query-language for care-pathway mining and analysis”, february 1st 2015, David Gross-Amblard, Thomas Guyet, André Happe

PhD in progress: Alban Siffer, “DataMining approaches for cyber attack detection”, mars 2016, Pierre-Alain Fouque, Alexandre Termier, Christine Largouët

10.2.3. Juries

Committee member of Rémy Dautriche Phd defense (Université de Grenoble Alpes): A. Termier

Reviewer of Olivier Cavadenti Phd (INSA Lyon): A. Termier

Committee member of Samir Loudni HDR (Université de Caen): A. Termier

Thesis advisory committee member of Jean Coquet (Univ. Rennes 1): A. Termier

Thesis advisory committee member of Benoit Bellot (INRA/IGEPP): T. Guyet

Thesis advisory committee member of Zhi Cheng (UNC/PPME): T. Guyet

10.3. Popularization

M.-O. Cordier is editorial board member of Interstices webzine.

11. Bibliography

Major publications by the team in recent years

[1] T. GUYET, R. QUINIOU. *Extracting temporal patterns from interval-based sequences*, in "International Join Conference on Artificial Intelligence (IJCAI)", Barcelone, Spain, July 2011, https://hal.inria.fr/inria-00618444

Publications of the year

Articles in International Peer-Reviewed Journals

[2] H. MIRISAEE, E. GAUSSIER, A. TERMIER. *Efficient local search for L_1 and L_2 binary matrix factorization*, in "Intelligent Data Analysis", 2016, vol. 20, pp. 783 - 807 [*DOI :* 10.3233/IDA-160832], https://hal.archives-ouvertes.fr/hal-01405186

International Conferences with Proceedings

[3] R. DAUTRICHE, A. TERMIER, R. BLANCH, M. SANTANA. *Towards Visualizing Hidden Structures*, in "International Conference on Data Mining (ICDM) / PhD Forum", Barcelone, Spain, 2016, https://hal.archives-ouvertes.fr/hal-01407664

[4] Y. DAUXAIS, D. GROSS-AMBLARD, T. GUYET, A. HAPPE. *Extraction de chroniques discriminantes*, in "Extraction et Gestion des Connaissances (EGC)", Grenoble, France, January 2017, https://hal.inria.fr/hal-01413473

[5] M. GEBSER, T. GUYET, R. QUINIOU, J. ROMERO, T. SCHAUB. *Knowledge-based Sequence Mining with ASP*, in "IJCAI 2016- 25th International joint conference on artificial intelligence", New-york, United States, AAAI, July 2016, 8 p. , https://hal.inria.fr/hal-01327363

[6] T. GUYET, Y. MOINARD, J. NICOLAS, R. QUINIOU. *Packing graphs with ASP for landscape simulation*, in "IJCAI 2016 - 25th International joint conference on artificial intelligence", New-york, United States, AAAI, July 2016, 8 p. , https://hal.inria.fr/hal-01327368

[7] C. LARGOUËT, Y.-M. BOZEC, Y. ZHAO. *Integrating socio-economic drivers in an explicit-time, qualitative fisheries model: EcoMata*, in "Understanding Marine Socio-ecological Systems including the human dimension in integrated ecosystem assessment", Brest, France, ICES, May 2016, https://hal.archives-ouvertes.fr/hal-01398963

[8] C. LARGOUËT, O. KRICHEN, Y. ZHAO. *Temporal Planning with extended Timed Automata*, in "28th International Conference on Tools with Artificial Intelligence (ICTAI 2016)", SAN JOSE, United States, N. BOURBAKIS, A. ESPOSITO, A. MALI, M. ALAMANIOTIS (editors), IEEE, November 2016, https://hal.archives-ouvertes.fr/hal-01398695

[9] C. LARGOUËT, Y. ZHAO, M.-O. CORDIER. *Searching for Cost-Optimized Strategies: An Agricultural Application*, in "2nd International Conference, ICDSST 2016", Plymouth, United Kingdom, Decision Support Systems VI - Addressing Sustainability and Societal Challenges Volume 250, Springer, May 2016, vol. Lecture Notes in Business Information Processing, n^{o} 250, pp. 31 - 43, Best Paper Award [*DOI :* 10.1007/978-3-319-32877-5_3], https://hal.archives-ouvertes.fr/hal-01398660

National Conferences with Proceedings

[10] T. GUYET, Y. MOINARD, R. QUINIOU, T. SCHAUB. *Fouille de motifs séquentiels avec ASP*, in "Extraction et Gestion de Connaissances (EGC)", Reims, France, Actes de la conférence Extraction et Gestion de Connaissances, RNTI, 2016, https://hal.inria.fr/hal-01239501

Conferences without Proceedings

[11] F. BALUSSON, M.-A. BOTREL, O. DAMERON, Y. DAUXAIS, E. DREZEN, A. DUPUY, T. GUYET, D. GROSS-AMBLARD, A. HAPPE, N. LE MEUR, B. LE NAUTOUT, E. LERAY, E. NOWAK, C. RAULT, E. OGER, E. POLARD. *PEPS: a platform for supporting studies in pharmaco-epidemiology using medico-administrative databases*, in "International Congress on e-Health Research", Paris, France, October 2016, https://hal.inria.fr/hal-01380939

[12] B. DOUX, C. GAUTRAIS, B. NEGREVERGNE. *Detecting strategic moves in HearthStone matches*, in "Machine Learning and Data Mining for Sports Analytics Workshop of ECML/PKDD", Riva del Garda, Italy, September 2016, https://hal.archives-ouvertes.fr/hal-01412432

[13] H.-S. PHAM, D. LAVENIER, A. TERMIER. *Identifying Genetic Variant Combinations using Skypatterns*, in "7th International Workshop on Biological Knowledge Discovery and Data Mining (Workshop BIOKDD '16)", Porto, Portugal, DEXA, September 2016 [*DOI :* 10.1109/DEXA.2016.13], https://hal.inria.fr/hal-01385614

[14] W. UGARTE, A. TERMIER, M. SANTANA. *Steady Patterns*, in "Data Science and Big Data Analytics workshop of International Conference on Data Mining", Barcelone, Spain, 2016, https://hal.archives-ouvertes.fr/hal-01408397

Scientific Books (or Scientific Book chapters)

[15] A. BAILLY, S. MALINOWSKI, R. TAVENARD, L. CHAPEL, T. GUYET. *Dense Bag-of-Temporal-SIFT-Words for Time Series Classification*, in "Advanced Analysis and Learning on Temporal Data", Springer, 2016, https://hal.archives-ouvertes.fr/hal-01252726

Other Publications

[16] C. GAUTRAIS, P. CELLIER, T. GUYET, R. QUINIOU, A. TERMIER. *Understanding Customer Attrition at an Individual Level: a New Model in Grocery Retail Context*, March 2016, International Conference on Extending Database Technology (EDBT), Poster [*DOI :* 10.5441/002/EDBT.2016.87], https://hal.archives-ouvertes.fr/hal-01405172

References in notes

[17] S. COLAS, C. COLLIN, P. PIRIOU, M. ZUREIK. *Association between total hip replacement characteristics and 3-year prosthetic survivorship: A population-based study*, in "JAMA Surgery", 2015, vol. 150, n^{o} 10, pp. 979–988

[18] G. MOULIS, M. LAPEYRE-MESTRE, A. PALMARO, G. PUGNET, J.-L. MONTASTRUC, L. SAILLER. *French health insurance databases: What interest for medical research?*, in "La Revue de Médecine Interne", 2015, vol. 36, n^{o} 6, pp. 411 - 417

[19] E. NOWAK, A. HAPPE, J. BOUGET, F. PAILLARD, C. VIGNEAU, P.-Y. SCARABIN, E. OGER. *Safety of Fixed Dose of Antihypertensive Drug Combinations Compared to (Single Pill) Free-Combinations: A Nested Matched Case–Control Analysis*, in "Medicine", 2015, vol. 94, n^{o} 49, e2229 p.

[20] E. POLARD, E. NOWAK, A. HAPPE, A. BIRABEN, E. OGER. *Brand name to generic substitution of antiepileptic drugs does not lead to seizure-related hospitalization: a population-based case-crossover study*, in "Pharmacoepidemiology and drug safety", 2015, vol. 24, n^{o} 11, pp. 1161–1169