IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2019

Project-Team ORPAILLEUR

Knowledge discovery, knowledge engineering

IN COLLABORATION WITH: Laboratoire lorrain de recherche en informatique et ses applications (LORIA)

RESEARCH CENTER
Nancy - Grand Est

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team ORPAILLEUR

Creation of the Project-Team: 2008 January 01

Keywords:

Computer Science and Digital Science:

A3. - Data and knowledge
A3.1. - Data
A3.1.1. - Modeling, representation
A3.1.7. - Open data
A3.2. - Knowledge
A3.2.1. - Knowledge bases
A3.2.2. - Knowledge extraction, cleaning
A3.2.3. - Inference
A3.2.4. - Semantic Web
A3.2.5. - Ontologies
A3.2.6. - Linked data
A3.3. - Data and knowledge analysis
A3.3.2. - Data mining
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.5. - Bayesian methods
A3.4.8. - Deep learning
A3.5.2. - Recommendation systems
A4.8. - Privacy-enhancing technologies
A8.1. - Discrete mathematics, combinatorics
A9.1. - Knowledge
A9.2. - Machine learning
A9.6. - Decision support
A9.8. - Reasoning

Other Research Topics and Application Domains:

B1.1. - Biology
B2. - Health
B2.3. - Epidemiology
B2.4.1. - Pharmaco kinetics and dynamics
B2.4.2. - Drug resistance
B3.1. - Sustainable development
B3.5. - Agronomy
B3.6. - Ecology
B3.6.1. - Biodiversity
B5. - Industry of the future
B9.5.6. - Data science

1. Team, Visitors, External Collaborators

Research Scientists

Amedeo Napoli [Team leader (until Oct 1st), CNRS, Senior Researcher, HDR]
Esther Catherine Galbrun [Inria, Researcher, on sabbatical leave since 2018]
Chedy Raïssi [Inria, Researcher, on sabbatical leave since 2019]

Faculty Members

Miguel Couceiro [Team leader (since Oct 1st), Univ. de Lorraine, Professor, HDR]
Adrien Coulet [Univ. de Lorraine, Associate Professor, sabbatical leave at University of Stanford, USA, until September 1st, HDR]
Sébastien Da Silva [Univ. de Lorraine, Associate Professor]
Jean-François Mari [Univ. de Lorraine, Professor, HDR]
Yannick Toussaint [Univ. de Lorraine, Professor, HDR]

Post-Doctoral Fellows

Alexandre Bazin [Univ de Lorraine]
Abdelkader Ouali [Inria (until August 31)]

PhD Students

Nacira Abbas [Inria]
Guilherme Alves Da Silva [Inria]
Quentin Brabant [Univ. de Lorraine, until Feb 1st]
Laurine Huber [Univ. de Lorraine]
Nyoman Juniarta [CNRS]
Tatiana Makhalova [Inria]
Pierre Monnin [Univ. de Lorraine]
François Pirot [Univ. de Lorraine]
Justine Reynaud [Univ. de Lorraine, ATER]
Claire Theobald [CNRS, from Dec 2019]
Laura Alejandra Zanella Calzada [Univ de Lorraine, from Nov 2019]
Georgios Zervakis [Inria, from Nov 2019]

Technical staff

Nicolas Dante [Univ de Lorraine, Engineer, from June 2019 until August 2019]
Laureline Nevin [Inria, Engineer]

Interns and Apprentices

Clement Bellanger [Univ de Lorraine, from Feb 2019 until Jun 2019]
Morgane Colle [Univ de Lorraine, from Feb 2019 until Aug 2019]
Romain Dalbard [Univ de Lorraine, from Jun 2019 until Aug 2019]
Victor Freyer [Univ de Lorraine, from Apr 2019 until Sep 2019]
Bofeng Huang [Univ de Lorraine, from Apr 2019 until Sep 2019]
Murat Kocak [Univ de Lorraine, from Jul 2019 until Oct 2019]
Melvin Moreau [Univ de Lorraine, from Jun 2019 until Aug 2019]
Maryam Naderan [Univ de Lorraine, from Mar until July 2019]
Gaurav Shajepal [Univ de Lorraine, from Mar 2019 until Sep 2019]
Yoann Simon [Univ de Lorraine, from Apr 2019 until May 2019]
Mayssaa Zeaiter [Inria, from Apr 2019 until Sep 2019]

Administrative Assistants

Emmanuelle Deschamps [Inria, Administrative Assistant]
Delphine Hubert [Univ. de Lorraine, Administrative Assistant]
Annick Jacquot [CNRS, Administrative Assistant]
Martine Kuhlmann [CNRS, Administrative Assistant (until Apr 2019)]
Messaoudi Anne-Marie [Univ de Lorraine, Administrative Assistant (since Apr 2019)]

External Collaborators

Alexandre Blansché [Univ. de Lorraine, Metz, Associate Professor]
Lydia Boudjeloud-Assala [Univ. de Lorraine, Metz, Associate Professor, HDR]
Brieuc Conan-Guez [Univ. de Lorraine, Metz, Associate Professor]
Alain Gély [Univ. de Lorraine, Metz, Associate Professor]
Florence Le Ber [ENGEES Strasbourg, Professor, HDR]
Frédéric Pennerath [Centrale-Supelec Metz, Associate Professor]

2. Overall Objectives

2.1. Introduction

Knowledge discovery in databases (KDD) consists in processing large volumes of data in order to discover knowledge units that are significant and reusable. Assimilating knowledge units to gold nuggets, and databases to lands or rivers to be explored, the KDD process can be likened to the process of searching for gold. This explains the name of the research team: in French "orpailleur" denotes a person who is searching for gold in rivers or mountains. The KDD process is based on three main operations: data preparation, data mining and interpretation of the extracted units as knowledge units. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the analyst. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. In this view, KDD is an exploratory process similar to "exploratory data analysis".

As a person searching for gold may have a certain experience about the task and the location, the analyst may use general and domain knowledge for improving the whole KDD process. Accordingly, the KDD process may be associated with knowledge bases –or domain ontologies– related to the domain of data for implementing *knowledge discovery guided by domain knowledge* (KDDK). In KDDK, extracted units may have "a life" after the interpretation step for becoming "actionable": they are represented as knowledge units using a knowledge representation formalism and integrated within an ontology to be reused for problem-solving needs. In this way, knowledge discovery extends and updates existing knowledge bases, materializing a complementarity between knowledge discovery and knowledge engineering.

3. Research Program

3.1. Hybrid and Exploratory Knowledge Discovery

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining, data exploration, formal concept analysis, classification, pattern mining, numerical methods in data mining.

Knowledge discovery in databases (KDD) aims at discovering intelligible and reusable patterns in possibly large databases. These patterns can then be interpreted as knowledge units to be reused in knowledge-based systems. From an operational point of view, the KDD process is based on three main steps: (i) selection and preparation of the data, (ii) data mining, (iii) interpretation of the discovered patterns. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the analyst. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. In this view, KDD is an exploratory process similar to "exploratory data analysis".

The KDD process –as implemented in the Orpailleur team– is based on data mining methods which are either symbolic or numerical. Symbolic methods are based on pattern mining (e.g. mining frequent itemsets, association rules, sequences...), Formal Concept Analysis (FCA) and extensions such as Pattern Structures and Relational Concept Analysis (RCA), and redescription mining. Numerical methods are based on Random Forests, Support Vector Machines (SVM), Neural Networks, and probabilistic approaches such as second-order Hidden Markov Models (HMM). Moreover, for being able to deal with complex data, numerical data mining methods can be associated with symbolic methods, for improving applicability and efficiency of knowledge discovery. This is particularly true in classification, where supervised and unsupervised approaches may be combined with benefits.

A main operation in the research work of Orpailleur is "classification", which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. In this way, domain knowledge, when available, can improve and guide the KDD process, materializing the idea of *Knowledge Discovery guided by Domain Knowledge* or KDDK. In KDDK, domain knowledge plays a role at each step of KDD: the discovered patterns can be interpreted as knowledge units and reused for problem-solving activities in knowledge systems, implementing the exploratory process "mining, interpreting, modeling, representing, and reasoning". Then knowledge discovery can be considered as a key task in knowledge engineering (KE), having an impact in various semantic activities, e.g. information retrieval, recommendation, and ontology engineering. In addition, if knowledge discovery can feed knowledge-based systems, in turn, domain knowledge can be used to support the knowledge discovery process.

Finally, life sciences, i.e. agronomy, biology, chemistry, and medicine, are application domains where the Orpailleur team has a very rich experience. The team intends to keep and to extend this experience, paying also more attention to the impact of knowledge discovery in the real world. This should lead to the design of green (sustainable), explainable, and fair data mining systems.

3.2. Text Mining

Keywords: text mining, knowledge discovery from texts, text classification, annotation, ontology engineering from texts.

The objective of a text mining process is to extract useful knowledge units from large collections of texts [71]. The text mining process shows specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making text mining a difficult task. A text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods.

From a knowledge discovery perspective, text mining aims at extracting "interesting units" (nouns and relations) from texts with the help of domain knowledge encoded within a knowledge base. The process is roughly similar for text annotation. Text mining is especially useful in the context of semantic web for ontology engineering. In the Orpailleur team, we work on the mining of real-world texts in application domains such as biology and medicine, using numerical and symbolic data mining methods. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.3. Knowledge Systems and Web of Data

Keywords: knowledge engineering, web of data, semantic web, ontology, description logics, classification-based reasoning, case-based reasoning, information retrieval, recommendation.

The web of data constitutes a good platform for experimenting ideas on knowledge engineering (KE) and knowledge discovery. A software agent may be able to read, understand, and manipulate information on the web, if and only if the knowledge necessary for achieving those tasks is available. This is why domain knowledge and ontologies are of main importance. OWL ("Web Ontology Language" https://www.w3.org/OWL/) is based on description logics (DLs [72]) and is the representation language commonly used for

designing ontologies. In OWL, knowledge units are represented by classes having properties and instances. Concepts are organized within a partially ordered set based on a subsumption relation, and the inference services are based on subsumption and classification.

Actually, there are many interconnections between concept lattices in FCA and ontologies, e.g. the partial order underlying an ontology can be supported by a concept lattice. Moreover, a pair of implications within a concept lattice can provide a possible materialization of a concept definition in an ontology. In this way, we study how the web of data, considered as a set of knowledge sources, e.g. DBpedia, Wikipedia, Yago, Freebase, can be mined for guiding the design of a knowledge base, and further, how knowledge discovery techniques can be applied for allowing a better usage of the web of data, e.g. Linked Open Data (LOD) classification and completion.

Then, a part of the research work in Knowledge Engineering is oriented towards knowledge discovery in the web of data, as, with the increased interest in machine processable data, more and more data is now published in RDF (Resource Description Framework) format. Particularly, we are interested in the completeness of the data and their potential to provide concept definitions in terms of necessary and sufficient conditions. We have proposed algorithms based on FCA and Redescription Mining which allow data exploration as well as the discovery of definition (bidirectional implication rules).

4. Application Domains

4.1. Life Sciences: Agronomy, Biology, Chemistry, and Medicine

Keywords: knowledge discovery in life sciences, biology, chemistry, medicine, pharmacogenomics and precision medicine.

One major application domain which is currently investigated by the Orpailleur team is related to life sciences, with particular emphasis on biology, medicine, and chemistry. The understanding of biological systems provides complex problems for computer scientists, and the developed solutions bring new research ideas or possibilities for biologists and for computer scientists as well. Indeed, the interactions between researchers in biology and researchers in computer science improve not only knowledge about systems in biology, chemistry, and medicine, but knowledge about computer science as well.

Knowledge discovery is gaining more and more interest and importance in life sciences for mining either homogeneous databases such as protein sequences and structures, or heterogeneous databases for discovering interactions between genes and the environment, or between genetic and phenotypic data, especially for public health and precision medicine (pharmacogenomics). Pharmacogenomics is one main challenge for the Orpailleur team as it considers a large panel of complex data ranging from biological to medical data, and various kinds of encoded domain knowledge ranging from texts to formal ontologies.

On the same line as biological data, chemical data are presenting important challenges w.r.t. knowledge discovery, for example for mining collections of molecular structures and collections of chemical reactions in organic chemistry. The mining of such collections is an important task for various reasons including the challenge of graph mining and the industrial needs (especially in drug design, pharmacology and toxicology). Molecules and chemical reactions are complex data that can be modeled as labeled graphs. Graph mining and Formal Concept Analysis methods play an important role in this application domain and can be used in an efficient and well-founded way [87].

Finally, research in agronomy is mainly based on cooperation between Inria and INRA. One research dimension is related to the characterization and the simulation of hedgerow structures in agricultural landscapes, based on Hilbert-Peano curves and Markov models [79]. Another research dimension is based on the mining of survey data for evaluating groundwater quality risks [86].

5. Highlights of the Year

5.1. Highlights of the Year

This year we would like to mention two publications as highlights of the year.

- The conference paper [10] got the best paper award at the International Conference on Formal Concept Analysis 2019 in Frankfurt, June 2019 (https://icfca2019.frankfurt-university.de/).
- Classical properties of functions such as associativity, although algebraically easy to read, are hard to meaningfully interpret. In [18], Miguel Couceiro and colleagues showed that associative and quasi-trivial operations that are non-decreasing are characterized in terms of total and weak orderings through the so-called single-peakedness property introduced in social choice theory by Duncan Black. This enabled visual interpretations of the above mentioned algebraic properties, and the enumeration of such operations led to several, previously unknown, integer sequences in Sloane's On-Line Encyclopedia of Integer Sequences (http://www.oeis.org), e.g., `A292932`, `A292933`, and `A292934`.

BEST PAPER AWARD:

[42]
J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Using Redescriptions and Formal Concept Analysis for Mining Definitions Linked Data*, in "ICFCA 2019 - 15th International Conference on Formal Concept Analysis", Francfort, Germany, June 2019, https://hal.inria.fr/hal-02170760

6. New Software and Platforms

6.1. ARPEnTAge

Analyse de Régularités dans les Paysages : Environnement, Territoires, Agronomie

KEYWORDS: Stochastic process - Hidden Markov Models

FUNCTIONAL DESCRIPTION: ARPEnTAge is a software based on stochastic models (HMM2 and Markov Field) for analyzing spatio-temporal data-bases. ARPEnTAge is built on top of the CarottAge system to fully take into account the spatial dimension of input sequences. It takes as input an array of discrete data in which the columns contain the annual land-uses and the rows are regularly spaced locations of the studied landscape. It performs a Time-Space clustering of a landscape based on its time dynamic Land Uses (LUS). Displaying tools and the generation of Time-dominant shape files have also been defined.

- Partner: INRA
- Contact: Jean-François Mari
- URL: http://carottage.loria.fr/index_in_english.html

6.2. CarottAge

KEYWORDS: Stochastic process - Hidden Markov Models

FUNCTIONAL DESCRIPTION: The system CarottAge is based on Hidden Markov Models of second order and provides a non supervised temporal clustering algorithm for data mining and a synthetic representation of temporal and spatial data. CarottAge is currently used by INRA researchers interested in mining the changes in territories related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination. CarottAge is also used for mining hydromorphological data. Actually a comparison was performed with three other algorithms classically used for the delineation of river continuum and CarottAge proved to give very interesting results for that purpose.

- Participants: Florence Le Ber and Jean-François Mari
- Partner: INRA
- Contact: Jean-François Mari
- URL: http://carottage.loria.fr/index_in_english.html

6.3. CORON

KEYWORDS: Data mining - Closed itemset - Frequent itemset - Generator - Association rule - Rare itemset

FUNCTIONAL DESCRIPTION: The Coron platform is a KDD toolkit organized around three main components: (1) Coron-base, (2) AssRuleX, and (3) pre- and post-processing modules.

The Coron-base component includes a complete collection of data mining algorithms for extracting itemsets such as frequent itemsets, closed itemsets, generators and rare itemsets. In this collection we can find APriori, Close, Pascal, Eclat, Charm, and, as well, original algorithms such as ZART, Snow, Touch, and Talky-G. AssRuleX generates different sets of association rules (from itemsets), such as minimal non-redundant association rules, generic basis, and informative basis. In addition, the Coron system supports the whole life-cycle of a data mining task and proposes modules for cleaning the input dataset, and for reducing its size if necessary.

- Participants: Adrien Coulet, Aleksey Buzmakov, Amedeo Napoli, Florent Marcuola, Jérémie Bourseau, Laszlo Szathmary, Mehdi Kaytoue, Victor Codocedo and Yannick Toussaint
- Contact: Amedeo Napoli
- URL: http://coron.loria.fr/site/index.php

6.4. LatViz: Visualization of Concept Lattices

- Contact: Amedeo Napoli
- URL: http://latviz.loria.fr/
- KEYWORDS: Formal Concept Analysis, Pattern Structures, Concept Lattice, Implications, Visualization

FUNCTIONAL DESCRIPTION.

LatViz is a tool allowing the construction, the display and the exploration of concept lattices. LatViz proposes some noticeable improvements over existing tools and introduces various functionalities focusing on interaction with experts, such as visualization of pattern structures for dealing with complex non-binary data, AOC-poset which is composed of the core elements of the lattice, concept annotations, filtering based on various criteria and a visualization of implications [75]. This way the user can effectively perform interactive exploratory knowledge discovery as often needed in knowledge engineering.

The LatViz platform can be associated with the Coron platform and extends its visualization capabilities (see http://coron.loria.fr). Recall that the Coron platform includes a complete collection of data mining algorithms for extracting itemsets and association rules.

6.5. OrphaMine: Data Mining Platform for Orphan Diseases

- Contact: Laureline Nevin
- URL: http://orphamine.inria.fr/
- KEYWORDS: Bioinformatics, data mining, biology, health, data visualization, drug development.

FUNCTIONAL DESCRIPTION.

The OrphaMine platform enables visualization, data integration and in-depth analytics in the domain of "orphan diseases", where data is extracted from the OrphaData ontology (http://www.orpha.net/consor/cgi-bin/index.php). At present, we aim at building a true collaborative portal that will serve different actors: (i) a general visualization of OrphaData data for physicians working, maintaining and developing this knowledge database about orphan diseases, (ii) the integration of analytics (data mining) algorithms developed by the different academic actors, (iii) the use of these algorithms to improve our general knowledge of rare diseases.

6.6. Siren: Interactive and Visual Redescription Mining

- Contact: Esther Catherine Galbrun
- URL: http://siren.gforge.inria.fr/main/
- KEYWORDS: Redescription mining, Interactivity, Visualization.

FUNCTIONAL DESCRIPTION.

Siren is a tool for interactive mining and visualization of redescriptions. Redescription mining aims to find distinct common characterizations of the same objects and, vice versa, to identify sets of objects that admit multiple shared descriptions. The goal is to provide domain experts with a tool allowing them to tackle their research questions using redescription mining. Merely being able to find redescriptions is not enough. The expert must also be able to understand the redescriptions found, adjust them to better match his domain knowledge and test alternative hypotheses with them, for instance. Thus, Siren allows mining redescriptions in an anytime fashion through efficient, distributed mining, to examine the results in various linked visualizations, to interact with the results either directly or via the visualizations, and to guide the mining algorithm toward specific redescriptions.

New features, such as a visualization of the contribution of individual literals in the queries and the simplification of queries as a post-processing, have been added to the tool.

7. New Results

7.1. Mining of Complex Data

Participants: Nacira Abbas, Guilherme Alves Da Silva, Alexandre Bazin, Alexandre Blansché, Lydia Boudjeloud-Assala, Quentin Brabant, Brieuc Conan-Guez, Miguel Couceiro, Adrien Coulet, Sébastien Da Silva, Alain Gély, Laurine Huber, Nyoman Juniarta, Florence Le Ber, Tatiana Makhalova, Jean-François Mari, Pierre Monnin, Amedeo Napoli, Laureline Nevin, Abdelkader Ouali, François Pirot, Frédéric Pennerath, Justine Reynaud, Claire Theobald, Yannick Toussaint, Laura Alejandra Zanella Calzada, Georgios Zervakis.

7.1.1. FCA and Variations: RCA, Pattern Structures, and Biclustering

Advances in data and knowledge engineering have emphasized the needs for pattern mining tools working on complex and possibly large data. FCA, which usually applies to binary data-tables, can be adapted to work on more complex data. In this way, we have contributed to some main extensions of FCA, namely Pattern Structures, Relational Concept Analysis and application of the "Minimum Description Length" (MDL) within FCA. Pattern Structures (PS [80], [85]) allow building a concept lattice from complex data, e.g. numbers, sequences, trees and graphs. Relational Concept Analysis (RCA [90]) is able to analyze objects described both by binary and relational attributes and can play an important role in text classification and text mining. Many developments were carried out in pattern mining and FCA for improving data mining algorithms and their applicability, and for solving some specific problems such as information retrieval, discovery of functional dependencies and biclustering.

We got several results in the discovery of approximate functional dependencies [29], the mining of RDF data, the visualization of the discovered patterns, and redescription mining. Moreover, based on Relational Concept Analysis, we worked also on the discovery and representation of n-ary relations in the framework of FCA [3]. In the same way, reusing ideas form subgroup discovery, we have initiated a whole line of research on the covering of the pattern spaces based on the "Minimum Description Length" (MDL) principle and we are working on the adaptation of MDL within the FCA framework [36] [7].

We are also working on designing hybrid mining methods, based on mining methods able to deal with symbolic and numerical data in parallel. In the context of the GEENAGE project, we are interested in the identification, in biomedical data, of biomarkers that are predictive of the development of diseases in the elderly population. Actually, the data are issued from a preceding study on metabolomic data for the detection of diabetes of type 2 [23]. The problem can be viewed as a classification problem where features which are predictive of a class should be identified. This leads us to study the notions of prediction and discrimination in classification problems. Combining numerical machine learning methods such as random forests, neural networks, and SVM, then multicriteria decision making methods (Pareto fronts), and pattern mining methods (including FCA), we developed a hybrid mining approach for selecting the features which are the most predictive and/or discriminant. Then the selected features are organized within a concept lattice to be presented to the analyst together with the reasons for their selection. The concept lattice makes more easy and natural the understanding of the feature selection. As such, this approach can also be seen as an explicable mining method, where the output includes the reasons for which features are selected in terms of prediction and discrimination.

In the framework of the CrossCult European Project about cultural heritage, we worked on the mining of visitor trajectories in a museum or a touristic site. We presented a theoretical and practical research work about the characterization of visitor trajectories and the mining of these trajectories as sequences [83], [84]. The mining process is based on two approaches in the framework of FCA. We focused on different types of sequences and more precisely on subsequences without any constraint and frequent contiguous subsequences. We also introduced a similarity measure allowing us to build a hierarchical classification which is used for interpretation and characterization of the trajectories. A natural extension of this research work on the characterization of trajectories is related to recommendation, i.e. based on an actual trajectory, how to recommend next items to be visited? Biclustering is a good candidate for designing recommendation methods and we especially worked on this topic this current year. In particular, we worked on several aspects of biclustering in the framework of FCA and we also tried to build a generic and unified framework from which several biclustering methods can be derived [34], [52].

7.1.2. Redescription Mining

Redescription mining is one of the pattern mining methods developed in the team. This method aims at finding distinct common characterizations of the same objects and, reciprocally, at identifying sets of objects having multiple shared descriptions [89]. This is motivated by the idea that in scientific investigations data oftentimes have different nature. For example, they might originate from distinct sources or be cast over separate terminologies.

In order to gain insight into the phenomenon of interest, a natural task is to identify the correspondences existing between these different aspects. A practical example in biology consists in finding geographical areas having two characterizations, one in terms of their climatic profile and one in terms of the occupying species. Discovering such redescriptions can contribute to better understand the influence of climate over species distribution. Besides biology, redescription mining can be applied in many concrete domains.

Following this way, we applied redescription mining for analyzing and mining RDF data in the web of data with the objective of discovering definitions of concepts and as well disjunctions (incompatibilities) of concepts, for completing knowledge bases in a semi-automated way [41] [10]. Redescription mining is well adapted to the task as a definition is naturally based on two sides of an equation, a left-hand side and a right-hand side.

7.1.3. Text Mining

The research work in text mining is mainly based on two ongoing PhD theses. The first research subject is related to the study of discourse and argumentation structures in a text based on tree mining and redescription mining [33], while the second research work is related to the mining of Pubmed abstracts about rare diseases. In the first research line, we investigate the similarities existing between discourse and argumentation structures by aligning subtrees in a corpus where texts are annotated. Contrasting related work, here we focus on the comparison of substructures within the text and not only the matching of relations. Based on data mining techniques such as tree mining and redescription mining, we are able to show that the structures

underlying discourse and argumentation can be (partially) aligned. There the annotations related to discourse and argumentation allow us to derive a mapping between the structures. In addition, the approach enables the study of similarities between diverse discourse structures, and as well the differences in terms of expressive power.

In the second research line, the objective is to discover features related to rare diseases, e.g. symptoms, related diseases, treatments, and possible disease evolution or variations. The texts to be analyzed are from Pubmed, i.e. a platform collecting millions of publications in the medical domain. This research project aims at developing new methods and tools for supporting knowledge discovery in textual data by combining methods from Natural Language Processing (NLP) and Knowledge Discovery in Databases (KDD). Here a key idea is to design an interacting and convergent process where NLP methods are used for guiding text mining and KDD methods are used for analyzing textual documents. In this way, NLP is based on extraction of general and temporal information, while KDD methods are especially based on pattern mining, FCA, and graph mining.

7.1.4. Consensus, Aggregation Functions and Multicriteria Decision Aiding Functions

Aggregation and consensus theory study processes dealing with the problem of merging or fusing several objects, e.g., numerical or qualitative data, preferences or other relational structures, into a single or several objects of similar type and that best represents them in some way. Such processes are modeled by so-called aggregation or consensus functions [81], [82]. The need to aggregate objects in a meaningful way appeared naturally in classical topics such as mathematics, statistics, physics and computer science, but it became increasingly emergent in applied areas such as social and decision sciences, artificial intelligence and machine learning, biology and medicine.

We are working on a theoretical basis of a unified theory of consensus and to set up a general machinery for the choice and use of aggregation functions. This choice depends on properties specified by users or decision makers, the nature of the objects to aggregate as well as computational limitations due to prohibitive algorithmic complexity. This problem demands an exhaustive study of aggregation functions that requires an axiomatic treatment and classification of aggregation procedures as well as a deep understanding of their structural behavior. It also requires a representation formalism for knowledge, in our case decision rules and methods for discovering them. Typical approaches include rough-set and FCA approaches, that we aim to extend in order to increase expressivity, applicability and readability of results. Applications of these efforts already appeared and further are expected in the context of three multidisciplinary projects, namely the “Fight Heart Failure” (research project with the Faculty of Medicine in Nancy), the European H2020 “CrossCult” project, and the “ISIPA” (Interpolation, Sugeno Integral, Proportional Analogy) project.

In the context of the project RHU “Fighting Heart Failure” (that aims to identify and describe relevant bio-profiles of patients suffering from heart failure) we are dealing with biomedical data, highly complex and heterogeneous, that include, among other, sociodemographical aspects, biological and clinical features, drugs taken by the patients, etc. One of our main challenges is to define relevant aggregation operators on this heterogeneous patient data that lead to a clustering of the patients. Each cluster should correspond to a bio-profile, i.e. a subgroup of patients sharing the same form of the disease and thus the same diagnosis and medical care strategy. We are working on ways for comparing and clustering patients, namely, by defining multidimensional similarity measures on this complex and heterogeneous biomedical data. To this end, we recently proposed a novel approach, that we named “unsupervised extremely randomized trees” (UET), that is inspired by the frameworks of unsupervised random forests (URF) and of extremely randomized trees (ET). The empirical study of UET showed that it outperforms existing methods (such as URF) in running time, while giving better clustering. However, UET was implemented for numerical data only, and this is a drawback when dealing with biomedical data.

To overcome this limitation we have recently proposed an adaptation of UET [63] that is agnostic to variable types –numerical, symbolic or both–, that is robust to noise, to correlated variables, and to monotone transformations, thus drastically limiting the need for preprocessing. In addition, this provides similarity measures for clustering purposes that show outperforming results compared to state-of-the-art clustering methodologies.

Also, motivated by current trends in graph clustering for applications in the semantic web, and community identification in computer and social networks, we recently proposed a novel graph clustering method, i.e. GraphTrees [61], that is based on random decision trees to compute pairwise dissimilarities between vertices in vertex-attributed graphs. Unlike existing methodologies, it applies directly to graphs whose vertex-attributes are heterogeneous without preprocessing, and with promising results in benchmark datasets that are competitive with best known methods.

In the context of the project ISIPA, we mainly focused on the utility-based preference model in which preferences are represented as an aggregation of preferences over different attributes, structured or not, both in the numerical and qualitative settings. In the latter case, the Sugeno integral is widely used in multiple criteria decision making and decision under uncertainty, for computing global evaluations of items based on local evaluations (utilities). The combination of a Sugeno integral with local utilities is called a Sugeno utility functional (SUF). A noteworthy property of SUFs is that they represent multi-threshold decision rules. However, not all sets of multi-threshold rules can be represented by a single SUF. We showed how to represent any set of multi-threshold rules as a combination of SUFs. Moreover, we studied their potential advantages as a compact representation of large sets of rules, as well as an intermediary step for extracting rules from empirical datasets [51]. We also proposed a novel method [58] for learning sets of decision rules that optimally fit the training data and that favors short rules over long ones. This is a competitive alternative to other methods for monotonic classification as in [78].

7.2. Knowledge Discovery in Healthcare and Life Sciences

Participants: Alexandre Bazin, Miguel Couceiro, Adrien Coulet, Sébastien Da Silva, Florence Le Ber, Jean-François Mari, Pierre Monnin, Amedeo Napoli, Abdelkader Ouali, Yannick Toussaint.

7.2.1. Ontology-based Clustering of Biological Data

Biomedical objects can be characterized by ontology annotations. For example, Gene Ontology annotations provide information on the functions of genes, while Human Phenotype Ontology (HPO) annotations provide information about phenotypes associated with diseases. It is usual to consider such annotations in the analysis of biomedical data, most of the time annotations from only one single ontology. However, complex objects such as diseases can be annotated at the same time w.r.t. different ontologies, making clear distinct dimensions. We are investigating how annotations from several ontologies may be cooperating in disease classification. In particular, we classified Genetic Intellectual Disabilities, on the basis of their HPO annotations and of Gene Ontology annotations of genes known for being responsible for these diseases [88]. We used clustering algorithms based on semantic similarities that enable us to compare sets of annotations. In particular, this experiment illustrates the fact that considering several ontologies provides better results in clustering, while selecting the best set of ontologies to combine is depending on the dataset and on the classification task. This study is still going on.

7.2.2. Validation of Pharmacogenomic Knowledge

State of the art knowledge in pharmacogenomics is heterogeneous w.r.t. validation. Some units of knowledge are well validated, observed on a large population and already used in clinical practice, while a large majority of this knowledge is lacking validation and reproducibility, mainly because of scarce observation. Accordingly, validating state of the art knowledge in pharmacogenomics by mining Electronic Health Records (EHRs) is one objective of the ANR project "PractiKPharma" initiated in 2016 (http://practikpharma.loria.fr/).

To carry out this validation, we define a minimal data schema for pharmacogenomic knowledge units (PGxO ontology), which is instantiated with data of different provenance (e.g. biomedical databases, literature and EHRs). The output of this instantiation is a (unique) knowledge graph called PGxLOD (https://pgxlod.loria.fr/). We defined and applied a set of so-called "reconciliation rules" that compare and align whenever possible knowledge units of different provenance [9]. The results of these rule applications are of particular interest since they highlight knowledge units defined in various data and knowledge sources. We are continuing this effort by studying how graph convolutional networks enable us to learn and then to compare the representation of n-ary relationships in the form of graph embeddings [39].

In addition, following our participation in the Biohackathon 2018 in Paris (https://2018.biohackathon-europe.org/), we firstly updated PGxLOD and improved its quality, completeness, and interconnection with other resources. Secondly we mined PGxLOD and searched for explanations about molecular mechanisms of adverse drug responses. Preliminary results where presented at the MedInfo Conference [59].

7.2.3. Mining Electronic Health Records

In the context of the Snowball Inria Associate Team, we studied the use of Electronic Health Records (EHRs) to predict at first prescription the need for a patient to be prescribed with a reduced drug dose [6]. We particularly focused on drugs whose dosage is known to be sensitive and variable. We used data from the Stanford Hospital to construct cohorts of patients that either did or did not need a dose change for each considered drug. After feature selection, we trained Random Forest models which successfully predict whether a new patient will or not require a dose change after being prescribed one of 23 drugs among 22 drug classes. Several of these drugs are related to clinical guidelines that recommend dose reduction exclusively in the case of adverse reaction. For these cases, a reduction in dosage may be considered as a surrogate for an adverse reaction, which our system could help to predict and to prevent.

In collaboration with Stanford University, we continued studying the development of predictive models from EHR data, in particular to evaluate the risk of atherosclerotic cardiovascular diseases (ASCVD). The evaluation of ASCVD risk is crucial for deciding upon the prescription of preventive therapies such as statins and others lipid lowering therapies. The prevalence of these diseases is depending on subgroups in a population, such as African-American and Asian people, which are indeed under-represented in cohorts that were used to fit the model currently used in clinics to evaluate the risk of ASCVD [25]. Due to such under-representation, biases are appearing in the evaluation of the risk when considering these different subgroups in the population. Then we proposed a method and a predictive model that controls, to some extent, the variability in the prediction of ASCVD when considering such "foreign" subgroups [40].

7.3. Knowledge Engineering and Web of Data

Participants: Nacira Abbas, Alexandre Bazin, Miguel Couceiro, Adrien Coulet, Florence Le Ber, Pierre Monnin, Amedeo Napoli, Justine Reynaud, Yannick Toussaint.

A first research topic in this axis relies on knowledge discovery in the web of data. This follows the increase of data published in RDF (Resource Description Framework) format and the interest in machine processable data. The quick growth of Linked Open Data (LOD) has led to challenging aspects regarding quality assessment and data exploration of the RDF triples that shape the LOD cloud. In the team, we are particularly interested in the completeness and the quality of data and their potential to provide concept definitions in terms of necessary and sufficient conditions [73], [74]. We have proposed a novel technique based on Formal Concept Analysis which classifies subsets of RDF data into a concept lattice. This allows data exploration as well as the discovery of implication rules which are used to automatically detect possible completions of RDF data and to provide definitions. Experiments on the DBpedia knowledge base show that this kind of approach is well-founded and effective [41] [10]. In addition, it should be noticed that this research work also involves redescription mining, showing the potential complementarity between definition mining and redescription mining.

The second topic in this axis is related to dependencies [77]. In the relational database model, functional dependencies (FDs) indicate a functional relation between sets of attributes: the values of a set of attributes are determined by the values of another set of attributes. FDs can be generalized into relational dependencies, also known as "link keys" in the web of data [76]. For example, link keys may identify the same book or article in different bibliographical data sources, where a link key is a statement of the form: $\{\langle \texttt{auteur}, \texttt{creator}\rangle, \langle \texttt{titre}, \texttt{title}\rangle\}$ $linkkey$ $\langle \texttt{Livre}, \texttt{Book}\rangle$ stating that whenever an instance of the class `Livre` has the same values for properties `auteur` and `titre` as an instance of class `Book` has for properties `creator` and `title`, then they denote the same entity. Such link keys are more complex than FDs in databases in several respects and they raise new problems to solve [2].

One main objective of this research work is to follow the lines initiated in recent papers [29], and to extend to link keys the characterization of FDs and of Similarity Dependencies within FCA and pattern structures. Indeed, this is one of the objective of the ANR ELKER project. Accordingly, one purpose is to extend the initial proposals based on FCA and to provide adapted implementations. This is part of the thesis work of Nacira Abbas initiated at the end of 2018 [26]. Moreover, we are currently investigating possible connections with Relational Concept Analysis and redescription mining. We would like to study the formulation of the discovery of link keys in reusing and extending some construction heuristics that were developed in redescription mining. Actually, redescription mining is a data mining technique which aims at constructing pairs of descriptions, i.e., pairs of logical statements, one for each of two datasets, such that their support sets, i.e., the sets of objects that satisfy each statements of a pair, respectively, are most similar, as measured for example by their Jaccard index.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. AGREV-3

Participant: Jean-François Mari.

The AGREV 3 project (for "Agriculture Environment Vittel") is part of "Agrivair" –a subsidiary of Nestlé Waters– in actions to protect the natural resources of natural mineral water. We used ARPEnTAge to mine survey data about the Vittel-Contrexéville territory, which is suspected of groundwater quality risks [8]. This allowed us to locate regions having the same behavior. In addition, this provided a more contrasted simulation by eliminating the influence of stable zones (forests, permanent grasslands) and a more precise definition of a "neutral" model.

8.1.2. Hydreos

Participants: Nicolas Dante, Jean-François Mari, Amedeo Napoli.

Hydreos is a state organization, so-called "Pôle de compétitivité", aimed at monitoring and evaluating the quality of water and its delivery (http://www.hydreos.fr/fr). Actually, data about water resources rely on many agronomic variables, including land use successions. The data to be analyzed are obtained by surveys or by satellite images and describe the land use at the level of the agricultural parcel. Then there is a search for detecting changes in land use and for correlating these changes to groundwater quality. Accordingly, one main challenge in our participation in Hydreos is to process and analyze space-time data for reaching a better understanding of the changes in the organization of a territory. The systems ARPEnTAge and CarottAge are used in this context, especially by agronomists of INRA (ASTER Mirecourt http://www6.nancy.inra.fr/sad-aster).

On other aspects, we tested new deep graph convolutional learning over data provided by the SEDIF "Syndicat des eaux d'Île-de-France" to predict the likelihood of water leaks in a network of pipes and compared it with a master thesis where spatial point process techniques were used (master thesis of Nicolas Dante, M2 IMSD Nancy).

8.1.3. The Smart Knowledge Discovery Project

Participants: Laureline Nevin, Amedeo Napoli.

The SKD project for "Smart Knowledge Discovery" aims at analyzing complex industrial data for troubleshooting and decision making, and is funded by "Grand Est Region". We are working on exploratory knowledge discovery with the Vize company, which is based in Nancy and specialized in visualization-based data mining. The data which are under study are provided by the Arcelor-Mittal Steel Company and are related to the monitoring of rolling mills. Data are complex time series and the problem is related to a so-called "predictive maintenance", or how to anticipate problems in the furnaces and avoid their stop. In this way, one main objective of SKD is to combine sequence mining and visualization tools for recognizing temperature problems in the furnaces, and thus preventing the occurrences of defects in the outputs of the rolling mills.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR

9.1.1.1. ANR ELKER (2017–2020)

Participants: Nacira Abbas, Miguel Couceiro, Amedeo Napoli.

The objectives of the ELKER ANR Research Project (https://project.inria.fr/elker/) are to study, formalize, and implement the search for link keys in RDF data [2]. Link keys generalize database keys in two independent directions, as firstly they deal with RDF data and secondly they apply across two relation datasets. In this project, we study the discovery of link keys and reasoning with link keys, being based on the FCA formalism. The ELKER project relies on the competencies of the Orpailleur Team in FCA and pattern structure algorithms, and also in partition pattern structures which are related to the discovery of functional dependencies. This project involves the EPI Orpailleur at Inria Nancy Grand Est, the EPI MOEX at Inria Grenoble Rhône Alpes, and LIASD at Université Paris 8.

9.1.1.2. ANR PractiKPharma (2016–2020)

Participants: Miguel Couceiro, Adrien Coulet, Pierre Monnin, Amedeo Napoli, Yannick Toussaint.

PractiKPharma for "Practice-based evidences for actioning Knowledge in Pharmacogenomics" is an ANR research project (http://practikpharma.loria.fr/) about the validation of domain knowledge in pharmacogenomics. Pharmacogenomics is interested in understanding how genomic variations related to patients have an impact on drug responses. While most of the available knowledge in pharmacogenomics –state of the art knowledge– lies in the biomedical literature, with various levels of validation, an originality of PractiKPharma is to use Electronic Health Records (EHRs) to constitute cohorts of patients where to discover knowledge units. Indeed, these cohorts are mined for discovering potential pharmacogenomics patterns to be then validated w.r.t. literature knowledge for becoming actionable knowledge units. More precisely, firstly we have to discover pharmacogenomic patterns from the literature, and secondly we should confirm or moderate the interpretation and validation of these units by mining EHRs. Comparing knowledge patterns extracted from the literature with facts extracted from EHRs is a complex task depending on the EHR language –the literature is in English whereas EHRs are in French– and on knowledge level, as EHRs represent observations at the patient level whereas the literature is related to sets of patients. The PractiKPharma involves three other laboratories, namely LIRMM in Montpellier, SSPIM in St-Etienne, and CRC in Paris.

9.1.1.3. ANR AstroDeep (2019–2022)

Participants: Miguel Couceiro, Amedeo Napoli, Claire Theobald.

Astronomical surveys planned for the coming years will produce data that present analysis challenges not only because of their scale (hundreds of petabytes), but also by the complexity of the measurement challenges on very deep images (for instance subpercent-level measurement of colors or shapes on blended objects). New machine learning techniques appear very promising: once trained, they are very efficient and excel at extracting features from complex images. In the AstroDeep project, we aim at developing such machine learning techniques that can be applied directly on complex images without going through the traditional steps of astronomical image processing, that lose information at each stage. The developed techniques will help to leverage the observation capabilities of future surveys (LSST, Euclid, and WFIRST), and will allow a joint analysis of data.

The AstroDeep ANR Project involves three labs, namely APC Paris ("Astroparticules et Cosmologie Paris"), the Orpailleur Team at Inria Nancy Grand Est/LORIA, and "Département d'Astrophysique CEA Saclay".

9.1.2. Inria Project Labs, Exploratory Research Actions, and Technological Development Actions

Participants: Guilherme Alves Da Silva, Alexandre Bazin, Miguel Couceiro, Nyoman Juniarta, Tatiana Makhalova, Amedeo Napoli, Laureline Nevin, Abdelkader Ouali, Claire Theobald, Georgios Zervakis.

HyAiAI (IPL 2019-2022) Recent progress in Machine Learning (ML) and especially in Deep Learning has made ML present and prominent in a wide range of applications. However, current and efficient ML approaches rely on complex numerical models. Then, the decisions which are proposed may be accurate but cannot be easily explained to the layman, especially in some cases where complex and human-oriented decisions should be made, e.g. to get a loan or not, to obtain a chosen enrollment at university. The objectives of the HyAIAI IPL are to study the problem of making ML methods interpretable. For that, we will design hybrid ML approaches that combine state of the art numerical models (e.g. neural networks) with explainable symbolic models (e.g. pattern mining). More precisely, one goal is to integrate high level domain constraints into ML models, to provide model designers information on ill-performing parts of the model, and to give the layman/practitioner understandable explanations on the results of the ML model.

The HyAIAI IPL project involves seven Inria Teams, namely Lacodam in Rennes (project leader), Magnet and SequeL in Lille, Multispeech and Orpailleur in Nancy, and TAU in Saclay.

Ordem (ADT 2019-2020) One of the outputs of the former Hybride ANR project was the Orphamine system which aims at information retrieval and diagnosis aid in the domain of "rare diseases". The Orphamine system is based on domain knowledge, and in particular on medical ontologies such as ORDO ("Orphanet Rare Diseases Ontology") and HPO ("Human Phenotype Ontology"). In this way, the objective of the "Ordem" ADT is to update Orphamine, in making the system more accessible and more open. This requires many developments for developing the connections with domain knowledge, graph mining methods for retrieving relevant units in knowledge graphs, actual visualization tools, pattern mining, statistical decision tools for decision making (in particular log-linear models), and as well text mining tools for analyzing expert queries and medical literature about rare diseases. Such developments are and will be carried out until the end of next year, for making the system robust and publicly accessible through a web interface.

HyGraMi (PRE Inria 2018-2020) Finally, the so called "projet de recherche exploratoire" (PRE) HyGraMi for "Hybrid Graph Mining for the Design of New Antibacterials" is about the fight against resistance of bacteria to antibiotics. The objective of HyGraMi is to design a hybrid data mining system for discovering new antibacterial agents. This system should rely on a combination of numeric and symbolic classifiers, that will be guided by expert domain knowledge. The analysis and classification of the chemical structures is based on an interaction between symbolic methods e.g. graph mining techniques, and numerical supervised classifiers based on exact and approximate matching. This year we work on a method based on tree decomposition for performing feature selection and improving data lining of such complex molecular structures [49].

9.2. European Initiatives

9.2.1. The H2020 CrossCult Project (2016-2019)

Participants: Miguel Couceiro, Nyoman Juniarta, Amedeo Napoli.

The H2020 CrossCult [1] project aims at making "reflective history" a reality in the European cultural context, by enabling the re-interpretation of European (hi)stories through cross-border interconnections among cultural digital resources, citizen viewpoints and physical venues. The project has two main goals, (i) to lower cultural EU barriers and create unique cross-border perspectives, by connecting existing digital historical resources and by creating new ones through public participation, (ii) to create long-lasting experiences of social learning and entertainment that will provide a better understanding and re-interpretation of European history. To achieve

[1] http://www.crosscult.eu/

this, CrossCult aims at connecting and combining existing digital cultural assets, at increasing integration, interaction, and reflection about European past and present history. CrossCult was implemented w.r.t. four real-world pilots including cities, museums, and cultural sites. The role of the Orpailleur Team, in conjunction with the LORIA Kiwi Team, was to work on data mining –actually sequence mining– and recommendation, with a focus on the mining visitor trajectories in a museum or a touristic site, and on the definition of a visitor profile in connection with domain knowledge.

The CrossCult project involved many teams, namely Luxembourg Institute for Science and Technology and Centre Virtuel de la Connaissance sur l'Europe (Luxembourg, leaders of the project), University College London (England), University of Malta (Malta), University of Peloponnese and Technological Educational Institute of Athens (Greece), Università degli Studi di Padova (Italy), University of Vigo (Spain), National Gallery (London, England), and GVAM Guìas Interactivas (Spain), and the Kiwi Team from LORIA together with the Orpailleur team.

9.3. International Initiatives

9.3.1. Inria International Labs

Inria@SiliconValley
Associate Team involved in the International Lab:

9.3.1.1. Snowball

Title: Discovering knowledge on drug response variability by mining electronic health records

International Partner (Institution - Laboratory - Researcher):

University of Stanford (United States) - Department of Medicine, Stanford Center for Biomedical Informatics Research (BMIR) - Nigam Shah

Start year: 2017

See also: http://snowball.loria.fr/

Snowball (2017-2019) is an Inria Associate Team and the continuation of the preceding Associate Team called Snowflake (2014-2016). The objective of Snowball is to study drug response variability through the lens of Electronic Health Records (EHRs). This is motivated by the fact that many factors, genetic as well as environmental, contribute to different responses from people to the same drug. The mining of EHRs can bring substantial elements for understanding and explaining drug response variability.

Accordingly the objectives of Snowball are to identify in EHR repositories groups of patients which are responding differently to similar treatments, and then to characterize these groups and predict patient drug sensitivity. These objectives are complementary to those of the PractiKPharma ANR project. Moreover, Adrien Coulet finished in September 2019 a two-years sabbatical stay in the lab of Nigam Shah at Stanford University initiated in September 2017 (and partly granted by an "Inria délégation").

9.3.2. Informal International Partners: Research Collaboration with HSE Moscow

Participants: Alexandre Bazin, Nacira Abbas, Guilherme Alves Da Silva, Miguel Couceiro, Nyoman Juniarta, Tatiana Makhalova, Amedeo Napoli, Justine Reynaud.

An ongoing collaboration involves the Orpailleur team and Sergei O. Kuznetsov at Higher School of Economics in Moscow (HSE). Amedeo Napoli visited HSE laboratory several times while Sergei O. Kuznetsov visits Inria Nancy Grand Est every year. The collaboration is materialized by the joint supervision of students (such as the thesis of Aleksey Buzmakov defended in 2015 and the ongoing thesis of Tatiana Makhalova), and the the organization of scientific events, as the workshop FCA4AI with seven editions between 2012 and 2019 (see http://www.fca4ai.hse.ru).

This year, we participated in the writing of common publications around the thesis work of Tatiana Makhalova and the organization of one main event, namely the seventh edition of the FCA4AI workshop in August 2019 at the IJCAI Conference which was held in Macao China.

10. Dissemination

10.1. Scientific Events Organization, General Chairs, Scientific Chairs

- Amedeo Napoli was the scientific co-chair with Sergei Kuznetsov of the track "General Topics of Data Analysis" at the AIST Conference held in Kazan Russia on July 17-19 2019 (8th International Conference on Analysis of Images, Social Networks, and Texts http://aistconf.org/ and http://aistconf.org/board/).
- Amedeo Napoli was the scientific co-chair with Sergei O. Kuznetsov (HSE Moscow) and Sebastian Rudolph (TU Dresden) of the seventh workshop FCA4AI "What can do FCA for Artificial Intelligence", which was co-located with the IJCAI Conference in Macao China, August 10 2019 (see http://www.fca4ai.hse.ru/).
- Miguel Couceiro and Amedeo Napoli were the general and scientific chairs of the 26ièmes Rencontres de la Société Francophone de Classification (SFC 2019) that were held on September 3-5 at Inria NGE/LORIA Nancy (see https://project.inria.fr/sfc2019/).

10.2. Scientific Animation

- The scientific animation in the Orpailleur team is based on the Team Seminar which is called the "Malotec" seminar (http://malotec.loria.fr/). The Malotec seminar is held in general twice a month and is used either for general presentations of members of the team or for invited presentations of external researchers.
- Members of the Orpailleur team are all involved, as members or as head persons, in various national research groups.
- The members of the Orpailleur team are involved in the organization of conferences and workshops, as members of conference program committees (AAAI, ECAI, ECML-PKDD, ESWC, ICCBR, ICDM, ICFCA, IJCAI, ISWC, KDD, SDM...), as members of editorial boards, and finally in the organization of journal special issues.

10.3. Teaching - Supervision - Juries

10.3.1. Teaching

- All the permanent members of the Orpailleur team are involved in teaching at all levels and mainly at Université de Lorraine. Actually, most of the members of the Orpailleur team are employed on "Université de Lorraine" positions.
- Responsability of the 2nd year of the NLP Master's program in the IDMC, Université de Lorraine.
- Local coordination of the European Erasmus Mundus Master's program LCT (Language and Communication Technologies).

  The LCT Master's program "Language and communication Technologies" (LCT) is designed to provide students with practice–oriented knowledge in computational and theoretical linguistics, natural language processing, and computer science, to meet the demands of industry and research in these rapidly growing areas. The LCT consortium includes 7 European Universities, i.e. Saarland, Lorraine, Trento, Malta, Groningen, Charles in Prague, Basque Country, and includes several partners, e.g., DFKI, IBM (Czech Rep.), VICOMTECH, Sony (Europe), IBM (Ireland), and Inria (France).

- Responsability in teaching courses about Artificial Intelligence and Knowledge-Based Systems at TELECOM Nancy, a engineer school for graduation in computer science at Université de Lorraine.

10.3.2. Supervision – Juries

- The members of the Orpailleur team are also involved in student supervision, at all university levels, from under-graduate until post-graduate students, engineers, PhD, postdoc students.
- Finally, the permanent members of the Orpailleur team are involved in HDR and thesis defenses, being thesis referees or thesis committee members.

11. Bibliography

Major publications by the team in recent years

[1] M. ANSDELL, Y. IOANNOU, H. OSBORN, M. SASDELLI, J. SMITH, D. CALDWELL, J. JENKINS, C. RAÏSSI, D. ANGERHAUSEN. *Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning*, in "The Astrophysical Journal Letters", December 2018, vol. 869, n° 1, L7 p. [*DOI :* 10.3847/2041-8213/AAF23B], https://hal.inria.fr/hal-01957950

[2] M. ATENCIA, J. DAVID, J. EUZENAT, A. NAPOLI, J. VIZZINI. *Link key candidate extraction with relational concept analysis*, in "Discrete Applied Mathematics", 2019, pp. 1-19 [*DOI :* 10.1016/J.DAM.2019.02.012], https://hal.archives-ouvertes.fr/hal-02196757

[3] A. BAZIN, J. CARBONNEL, M. HUCHARD, G. KAHN, P. KEIP, A. OUZERDINE. *On-demand Relational Concept Analysis*, in "ICFCA: 15th International Conference on Formal Concept Analysis", Frankfurt, Germany, D. CRISTEA, F. L. BER, B. SERTKAYA (editors), Formal Concept Analysis, Springer International Publishing, 2019, vol. 11511, pp. 155-172 [*DOI :* 10.1007/978-3-030-21462-3_11], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02092140

[4] M. COUCEIRO, N. HUG, H. PRADE, G. RICHARD. *Behavior of Analogical Inference w.r.t. Boolean Functions*, in "IJCAI 2018 - 27th International Joint Conference on Artificial Intelligence", Stockholm, Sweden, July 2018, pp. 2057–2063, https://hal.inria.fr/hal-02139765

[5] M. COUCEIRO, E. LEHTONEN, P. MERCURIALI, R. PÉCHOUX. *On the efficiency of normal form systems for representing Boolean functions*, in "Theoretical Computer Science", 2019, forthcoming, https://hal.inria.fr/hal-02153506

[6] A. COULET, N. H. SHAH, M. WACK, M. CHAWKI, N. JAY, M. DUMONTIER. *Predicting the need for a reduced drug dose, at first prescription*, in "Scientific Reports", October 2018, vol. 8, n° 1 [*DOI :* 10.1038/S41598-018-33980-0], https://hal.inria.fr/hal-01901566

[7] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *Numerical Pattern Mining Through Compression*, in "DCC 2019 - 2019 Data Compression Conference", Snowbird, United States, IEEE, March 2019, pp. 112-121, https://hal.archives-ouvertes.fr/hal-02162927

[8] J.-F. MARI, A. GOBILLOT, M. BENOÎT. *Time Space Simulation of Land Use changes by stochastic modeling*, in "Revue Internationale de Géomatique", August 2018, vol. 28, n° 2, pp. 219–242, https://hal.inria.fr/hal-01662140

[9] P. MONNIN, J. LEGRAND, G. HUSSON, P. RINGOT, A. TCHECHMEDJIEV, C. JONQUET, A. NAPOLI, A. COULET. *PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison*, in "BMC Bioinformatics", April 2019, vol. 20, n° S4 [*DOI :* 10.1186/s12859-019-2693-9], https://hal.inria.fr/hal-02103899

[10] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Using Redescriptions and Formal Concept Analysis for Mining Definitions Linked Data*, in "ICFCA 2019 - 15th International Conference on Formal Concept Analysis", Francfort, Germany, June 2019, https://hal.inria.fr/hal-02170760

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] Q. BRABANT. *Lattice polynomial functions for interpolation and monotonic classification*, Université de Lorraine, January 2019, https://hal.univ-lorraine.fr/tel-02096400

[12] A. COULET. *Mappings between data, texts and knowledge for biomedical knowledge discovery*, Université de Lorraine, December 2019, Habilitation à diriger des recherches, https://hal.inria.fr/tel-02429926

[13] N. JUNIARTA. *Mining complex data and biclustering using formal concept analysis*, Université de Lorraine, December 2019, https://hal.inria.fr/tel-02426034

[14] J. REYNAUD. *Mining definitions in the web of data*, Université de Lorraine (Nancy), December 2019, https://hal.inria.fr/tel-02426421

Articles in International Peer-Reviewed Journals

[15] M. ATENCIA, J. DAVID, J. EUZENAT, A. NAPOLI, J. VIZZINI. *Link key candidate extraction with relational concept analysis*, in "Discrete Applied Mathematics", 2019, pp. 1-19 [*DOI :* 10.1016/J.DAM.2019.02.012], https://hal.archives-ouvertes.fr/hal-02196757

[16] Q. BRABANT, M. COUCEIRO, J. R. FIGUEIRA. *Interpolation by lattice polynomial functions: a polynomial time algorithm*, in "Fuzzy Sets and Systems", August 2019, vol. 368, pp. 101-118 [*DOI :* 10.1016/J.FSS.2018.12.009], https://hal.archives-ouvertes.fr/hal-01958903

[17] M. COUCEIRO, J. DEVILLET. *Every quasitrivial n-ary semigroup is reducible to a semigroup*, in "Algebra Universalis", 2019, forthcoming [*DOI :* 10.1007/s00012-019-0626-0], https://hal.inria.fr/hal-02099236

[18] M. COUCEIRO, J. DEVILLET, J.-L. MARICHAL. *Quasitrivial semigroups: Characterizations and enumerations*, in "Semigroup Forum", June 2019, vol. 98, n° 3, pp. 472-498 [*DOI :* 10.1007/s00233-018-9928-3], https://hal.inria.fr/hal-01826868

[19] M. COUCEIRO, E. LEHTONEN, P. MERCURIALI, R. PÉCHOUX. *On the efficiency of normal form systems for representing Boolean functions*, in "Theoretical Computer Science", 2019, forthcoming, https://hal.inria.fr/hal-02153506

[20] M. COUCEIRO, M. MARÓTI, T. WALDHAUSER, L. ZADORI. *Computing version spaces in the qualitative approach to multicriteria decision aid*, in "International Journal of Foundations of Computer Science", February 2019, vol. 30, n° 2, pp. 333-353 [*DOI :* 10.1142/S0129054119500084], https://hal.inria.fr/hal-01404590

[21] M. COUCEIRO, P. MERCURIALI, R. PÉCHOUX, A. SAFFIDINE. *On the complexity of minimizing median normal forms of monotone Boolean functions and lattice polynomials*, in "Journal of Multiple-Valued Logic and Soft Computing", 2019, vol. 33, n° 3, pp. 197-218, forthcoming, https://hal.inria.fr/hal-01905491

[22] E. FIMMEL, C. C. MICHEL, F. PIROT, J.-S. SERENI, L. STRÜNGMANN. *Mixed circular codes*, in "Mathematical Biosciences", July 2019, vol. 317, 108231 p. [*DOI :* 10.1016/J.MBS.2019.108231], https://hal.archives-ouvertes.fr/hal-02188407

[23] D. GRISSA, B. COMTE, M. PETERA, E. PUJOS-GUILLOT, A. NAPOLI. *A hybrid and exploratory approach to knowledge discovery in metabolomic data*, in "Discrete Applied Mathematics", January 2019, forthcoming [*DOI :* 10.1016/J.DAM.2018.11.025], https://hal.inria.fr/hal-02195463

[24] P. MONNIN, J. LEGRAND, G. HUSSON, P. RINGOT, A. TCHECHMEDJIEV, C. JONQUET, A. NAPOLI, A. COULET. *PGxO and PGxLOD: a reconciliation of pharmacogenomic knowledge of various provenances, enabling further comparison*, in "BMC Bioinformatics", April 2019, vol. 20, n° S4 [*DOI :* 10.1186/S12859-019-2693-9], https://hal.inria.fr/hal-02103899

[25] F. RODRIGUEZ, S. CHUNG, M. BLUM, A. COULET, S. BASU, L. PALANIAPPAN. *Atherosclerotic Cardiovascular Disease Risk Prediction in Disaggregated Asian and Hispanic Subgroups Using Electronic Health Records*, in "Journal of the American Heart Association", July 2019, vol. 8, n° 14 [*DOI :* 10.1161/JAHA.118.011874], https://hal.inria.fr/hal-02196129

International Conferences with Proceedings

[26] N. ABBAS, J. DAVID, A. NAPOLI. *Linkex: A Tool for Link Key Discovery Based on Pattern Structures*, in "ICFCA 2019 - workshop on Applications and tools of formal concept analysis", Frankfurt, Germany, Proc. ICFCA workshop on Applications and tools of formal concept analysis, 2019, pp. 33-38, abbas2019a, https://hal.archives-ouvertes.fr/hal-02168775

[27] G. ALVES, M. COUCEIRO, A. NAPOLI. *Sélection de mesures de similarité pour les données catégorielles*, in "20ème édition de la conférence Extraction et Gestion des Connaissances (EGC)", Bruxelles, Belgium, January 2020, https://hal.archives-ouvertes.fr/hal-02410221

[28] A. BAZIN, J. CARBONNEL, M. HUCHARD, G. KAHN, P. KEIP, A. OUZERDINE. *On-demand Relational Concept Analysis*, in "ICFCA: 15th International Conference on Formal Concept Analysis", Frankfurt, Germany, D. CRISTEA, F. LE BER, B. SERTKAYA (editors), Formal Concept Analysis, Springer International Publishing, 2019, vol. 11511, pp. 155-172 [*DOI :* 10.1007/978-3-030-21462-3_11], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02092140

[29] V. CODOCEDO, J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Sampling Representation Contexts with Attribute Exploration*, in "15th International Conference on Formal Concept Analysis", Frankfurt, Germany, D. CRISTEA, F. LE BER, B. SERTKAYA (editors), Proceedings of the 15th International Conference on Formal Concept Analysis (ICFCA 2019), Springer, May 2019, vol. Lecture Notes in Artificial Intelligence, n° 11511, pp. 307-314 [*DOI :* 10.1007/978-3-030-21462-3_20], https://hal.inria.fr/hal-02195498

[30] M. COUCEIRO, L. HADDAD, V. LAGERKVIST. *Fine-Grained Complexity of Constraint Satisfaction Problems through Partial Polymorphisms: A Survey (Dedicated to the memory of Professor Ivo Rosenberg)*, in "ISMVL2019 - IEEE 49th International Symposium on Multiple-Valued Logic", Fredericton, NB, Canada, May 2019, https://hal.inria.fr/hal-02190089

[31] M. COUCEIRO, L. HADDAD, M. POUZET. *The mathematics of Ivo Rosenberg (Dedicated to the memory of Professor Ivo Rosenberg)*, in "ISMVL2019 - IEEE 49th International Symposium on Multiple-Valued Logic", Fredericton, NB, Canada, May 2019, pp. 43-48, https://hal.inria.fr/hal-02190088

[32] M. COUCEIRO, A. NAPOLI. *Elements About Exploratory, Knowledge-Based, Hybrid, and Explainable Knowledge Discovery*, in "ICFCA 2019 - 15th International Conference on Formal Concept Analysis", Frankfurt, Germany, D. CRISTEA, F. LE BER, B. SERTKAYA (editors), Proceedings of the 15th International Conference on Formal Concept Analysis, Springer, 2019, vol. Lecture Notes in Artificial Intelligence, n° 11511, pp. 3-16 [*DOI :* 10.1007/978-3-030-21462-3_1], https://hal.inria.fr/hal-02195480

[33] L. HUBER, Y. TOUSSAINT, C. ROZE, M. DARGNAT, C. BRAUD. *Aligning Discourse and Argumentation Structures using Subtrees and Redescription Mining*, in "6th International Workshop on Argument Mining", Florence, Italy, Proceedings of the 6th Workshop on Argument Mining, August 2019, https://hal.archives-ouvertes.fr/hal-02165048

[34] N. JUNIARTA, M. COUCEIRO, A. NAPOLI. *A Unified Approach to Biclustering Based on Formal Concept Analysis and Interval Pattern Structures*, in "DS 2019 - 22nd International Conference on Discovery Science", Split, Croatia, Discovery Science - 22nd International Conference, October 2019, https://hal.inria.fr/hal-02266200

[35] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *Numerical Pattern Mining Through Compression*, in "DCC 2019 - 2019 Data Compression Conference", Snowbird, United States, IEEE, March 2019, pp. 112-121, https://hal.archives-ouvertes.fr/hal-02162927

[36] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *On Coupling FCA and MDL in Pattern Mining*, in "The 15th International Conference on Formal Concept Analysis", Frankfurt, Germany, D. CRISTEA, F. LE BER, B. SERTKAYA (editors), Springer, May 2019, vol. 11511, pp. 332-340, https://hal.archives-ouvertes.fr/hal-02162928

[37] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *Pattern Mining through compression: towards to probabilistic models*, in "Proceedings of the 17th Russian Conference on Artificial Intelligence", Ulyanovsk, Russia, O. P. KUZNETSOV, I. A. SOKOLOV, S. N. VASILIEV, G. S. OSIPOV (editors), Proceedings of the 17th Russian Conference on Artificial Intelligence, Yarushkina, Nadezhda G., October 2019, vol. 2, pp. 164-172, https://hal.archives-ouvertes.fr/hal-02192794

[38] T. MAKHALOVA, M. TRNECKA. *A Study of Boolean Matrix Factorization Under Supervised Settings*, in "ICFCA 2019 - The 15th International Conference on Formal Concept Analysis", Frankfurt, Germany, Springer, May 2019, pp. 341-348 [*DOI :* 10.1007/978-3-030-21462-3_24], https://hal.archives-ouvertes.fr/hal-02162929

[39] P. MONNIN, C. RAÏSSI, A. NAPOLI, A. COULET. *Knowledge Reconciliation with Graph Convolutional Networks: Preliminary Results*, in "DL4KG2019 - Workshop on Deep Learning for Knowledge Graphs", Portoroz, Slovenia, M. ALAM, D. BUSCALDI, M. COCHEZ, F. OSBORNE, D. R. RECUPERO, H. SACK (editors), June 2019, vol. CEUR Workshop Proceedings, n° 2377, https://hal.inria.fr/hal-02155546

[40] S. PFOHL, B. MARAFINO, A. COULET, F. RODRIGUEZ, L. PALANIAPPAN, N. SHAH. *Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk*, in "AIES '19 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society", Honolulu, United States, ACM Press, January 2019, pp. 271-278 [*DOI :* 10.1145/3306618.3314278], https://hal.inria.fr/hal-02388730

[41] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Redescription mining for learning definitions and disjointness axioms in Linked Open Data*, in "ICCS 2019 - 24th International Conference on Conceptual Structures", Marburg, Germany, July 2019, https://hal.inria.fr/hal-02170763

[42] *Best Paper*
J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Using Redescriptions and Formal Concept Analysis for Mining Definitions Linked Data*, in "ICFCA 2019 - 15th International Conference on Formal Concept Analysis", Francfort, Germany, June 2019, https://hal.inria.fr/hal-02170760.

National Conferences with Proceedings

[43] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Les forêts d'arbres extrêmement aléatoires : utilisation dans un cadre non supervisé*, in "EGC 2019 - 19ème Conférence Francophone sur l'Extraction et Gestion des connaissances", Metz, France, Hermann-Éditions, January 2019, vol. RNTI E-35, pp. 395-400, https://hal.inria.fr/hal-02099532

[44] L. HUBER, Y. TOUSSAINT, C. ROZE, M. DARGNAT, C. BRAUD. *Aligning Discourse and Argumentation Structures using Subtrees and Redescription Mining*, in "26èmes Rencontres de la Société Francophone de Classification (SFC)", Nancy, France, Actes des 26èmes Rencontres de la Société Francophone de Classification (SFC), September 2019, https://hal.archives-ouvertes.fr/hal-02266623

[45] N. JUNIARTA, M. COUCEIRO, A. NAPOLI. *Application des Pattern Structures à la découverte de biclusters à changements de signes cohérents*, in "EGC 2019 - 19ème Conférence francophone sur Extraction et Gestion des connaissances", Metz, France, Hermann-Éditions, January 2019, vol. RNTI E-35, pp. 285-290, https://hal.inria.fr/hal-02099607

[46] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *On Entropy in Pattern Mining*, in "SFC 2019 - XXVIe Rencontres de la Société Francophone de Classification", Nancy, France, September 2019, https://hal.archives-ouvertes.fr/hal-02193296

Conferences without Proceedings

[47] G. ALVES, M. COUCEIRO, A. NAPOLI. *Towards a Constrained Clustering Algorithm Selection*, in "26èmes Rencontres de la Société Francophone de Classification", Nancy, France, M. COUCEIRO, A. NAPOLI (editors), September 2019, vol. Actes des 26èmes Rencontres de la Société Francophone de Classification, https://hal.archives-ouvertes.fr/hal-02397436

[48] T. MAKHALOVA, S. O. KUZNETSOV, A. NAPOLI. *On compression, learning & searching regularity in big data*, in "13es journées scientifiques", Toulon, France, March 2019, https://hal.archives-ouvertes.fr/hal-02162931

[49] A. OUALI, N. JUNIARTA, B. MAIGRET, A. NAPOLI. *A Feature Selection Method based on Tree Decomposition of Correlation Graph*, in "LEG@ECML-PKDD 2019 - The third International Workshop on Advances in Managing and Mining Large Evolving Graphs", Würzburg, Germany, September 2019, https://hal.archives-ouvertes.fr/hal-02194229

Scientific Books (or Scientific Book chapters)

[50] D. ALLOUCHE, S. BARBE, S. DE GIVRY, G. KATSIRELOS, Y. LEBBAH, S. LOUDNI, A. OUALI, T. SCHIEX, D. SIMONCINI, M. ZYTNICKI. *Cost Function Networks to Solve Large Computational Protein Design Problems*, in "Operations Research and Simulation in healthcare", M. MASMOUDI, B. JARBOUI, P. SIARRY (editors), Springer, 2019, forthcoming, https://hal.archives-ouvertes.fr/hal-02177634

[51] M. COUCEIRO, D. DUBOIS, H. FARGIER, M. GRABISCH, H. PRADE, A. RICO. *New directions in ordinal evaluation: Sugeno integrals and beyond*, in "New Perspectives in Multiple Criteria Decision Making", M. DOUMPOS, J. FIGUEIRA, S. GRECO, C. ZOPOUNIDIS (editors), Springer, Cham, April 2019, https://hal.inria.fr/hal-01941776

[52] N. JUNIARTA, M. COUCEIRO, A. NAPOLI. *Order-preserving Biclustering Based on FCA and Pattern Structures*, in "Complex Pattern Mining: New Challenges, Methods and Applications", A. APPICE, M. CECI, C. LOGLISCI, G. MANCO, E. MASCIARI, Z. W. RAS (editors), Springer Series on Studies in Computational Intelligence, 2019, forthcoming, https://hal.inria.fr/hal-02181585

Books or Proceedings Editing

[53] M. COUCEIRO, A. NAPOLI (editors). *Société Francophone de Classification (SFC) Actes des 26èmes Rencontres*, Actes des 26èmes Rencontres de la Société Francophone de Classification (SFC), 2019, 147 p. , https://hal.inria.fr/hal-02432406

[54] S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors). *Workshop Notes of the Seventh International Workshop "What can FCA do for Artificial Intelligence?"*, CEUR-WS.org, 2019, vol. CEUR Workshop Proceedings 2529, 87 p. , https://hal.inria.fr/hal-02431335

[55] W. M. VAN DER AALST, V. BATAGELJ, D. I. IGNATOV, M. KHACHAY, V. KUSKOVA, A. KUTUZOV, S. O. KUZNETSOV, I. A. LOMAZOVA, N. LOUKACHEVITCH, A. NAPOLI, P. M. PARDALOS, M. PELILLO, A. V. SAVCHENKO, E. TUTUBALINA. , W. M. VAN DER AALST, V. BATAGELJ, I. A. LOMAZOVA, N. LOUKACHEVITCH, A. NAPOLI, P. M. PARDALOS, M. PELILLO, A. V. SAVCHENKO, E. TUTUBALINA, D. I. IGNATOV, M. KHACHAY, M. KHACHAY, V. KUSKOVA, A. KUTUZOV, S. O. KUZNETSOV (editors) *Analysis of Images, Social Networks and Texts*, Lecture Notes in Computer Science, Springer, 2019, vol. 11832, 426 p. [*DOI :* 10.1007/978-3-030-37334-4], https://hal.inria.fr/hal-02432920

Other Publications

[56] M. ALAM, T. A. GHORFI, E. AGRAWAL, O. ALQAWASMEH, A. ANNANE, C. D'AMATO, A. AZZAM, A. BEREZOVSKYI, R. BISWAS, M. BONDUEL, Q. BRABANT, C.-I. BUCUR, E. CAMOSSI, V. A. CARRIERO, S. CHARI, D. C. FRAGA, F. CIROKU, M. COCHEZ, V. CUTRONA, R. DANDAN, P. D. P. JIMNEZ, D. DESS, V. DI CARLO, A. E. A. DJEBRI, M. VAN ERP, F. M. FALAKH, A. F. IZQUIERDO, G. FUTIA, A. GANGEMI, S. GASPERONI, A. GRALL, L. HELING, P.-H. PARIS, N. HERRADI, S. ISSA, S. JOZASHOORI, N. JUNIARTA, L.-A. KAFFEE, I. KELES, P. KHARE, V. KOVTUN, V. LEONE, S. LI, S. LIEBER, P. LISENA, T. MAKHALOVA, L. MARINUCCI, T. MINIER, B. MOREAU, A. M. LOUSTAUNAU, D. NANDINI, S. OZDOWSKA, A. P. DE MOURA, S. PADHEE, G. PALMA, V. PRESUTTI, R. REDA, E. RIZZA, H. ROSALES-MNDEZ, S. RUDOLPH, H. SACK, L. SCIULLO, H. SIMANJUNTAK, C. STOMEO, T. THANAPALASINGAM, T. TIETZ, D. VARANKA, M.-E. VIDAL, M. WOLOWYK, M. ZOCHOLL. *Linked Open Data Validity – A Technical Report from ISWS 2018*, April 2019, https://arxiv.org/abs/1903.12554 - working paper or preprint, https://hal.inria.fr/hal-02087112

[57] G. ALVES, M. COUCEIRO, A. NAPOLI. *Similarity Measure Selection for Categorical Data Clustering*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02399640

[58] Q. BRABANT, M. COUCEIRO, D. DUBOIS, H. PRADE, A. RICO. *Learning rule sets and Sugeno integrals for monotonic classification problems*, December 2019, working paper or preprint, https://hal.inria.fr/hal-02427608

[59] F.-É. CALVIER, P. MONNIN, M. BOLAND, P. JARNOT, E. BRESSO, M. SMAÏL-TABBONE, A. COULET, C. BOUSQUET. *Providing Molecular Characterization for Unexplained Adverse Drug Reactions : Podium Abstract*, July 2019, Podium Abstract at MedInfo 2019, Lyon, France, https://hal.inria.fr/hal-02196134

[60] M. COUCEIRO, J. DEVILLET, J.-L. MARICHAL, P. MATHONET. *Reducibity of n-ary semigroups: from quasitriviality towards idempotency*, September 2019, working paper or preprint, https://hal.inria.fr/hal-02293908

[61] K. DALLEAU, M. COUCEIRO, M. SMAIL-TABBONE. *Computing Vertex-Vertex Dissimilarities Using Random Trees: Application to Clustering in Graphs*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02427563

[62] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Clustering graphs using random trees*, September 2019, working paper or preprint, https://hal.inria.fr/hal-02282207

[63] K. DALLEAU, M. COUCEIRO, M. SMAÏL-TABBONE. *Unsupervised Extra Trees: a stochastic approach to compute similarities in heterogeneous data*, January 2019, working paper or preprint, https://hal.inria.fr/hal-01982232

[64] E. FIMMEL, C. C. MICHEL, F. PIROT, J.-S. SERENI, L. STRÜNGMANN. *Comma-free Codes Over Finite Alphabets*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02376793

[65] N. JUNIARTA, V. CODOCEDO, M. COUCEIRO, M. KAYTOUE, A. NAPOLI. , D. CRISTEA, F. LE BER, R. MISSAOUI, L. KWUIDA, B. SERTKAYA (editors) *Pattern Structures for Identifying Biclusters with Coherent Sign Changes*, Proceedings of the 15th International Conference on Formal Concept Analysis (Supplementary Proceedings), June 2019, ICFCA 2019 - 15th International Conference on Formal Concept Analysis, https://hal.inria.fr/hal-02166713

[66] N. JUNIARTA, M. COUCEIRO, A. NAPOLI. *Formal Concept Analysis for Identifying Biclusters with Coherent Sign Changes*, July 2019, working paper or preprint, https://hal.inria.fr/hal-02181600

[67] P. MONNIN. *Discovering and Comparing Relational Knowledge, the Example of Pharmacogenomics*, January 2019, Article in Proceedings of the EKAW Doctoral Consortium 2018 co-located with the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018), https://hal.inria.fr/hal-01955424

[68] F. PIROT, J.-S. SERENI. *Fractional chromatic number, maximum degree and girth*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02096426

[69] A. ABBOUD, A. LAHMADI, M. RUSINOWITCH, M. COUCEIRO, A. BOUHOULA. *Minimizing Range Rules for Packet Filtering Using a Double Mask Representation*, May 2019, IFIP Networking 2019, Poster, https://hal.inria.fr/hal-02393008

[70] A. ABBOUD, A. LAHMADI, M. RUSINOWITCH, M. COUCEIRO, A. BOUHOULA, S. E. H. AWAINIA, M. AYADI. *Minimizing Range Rules for Packet Filtering Using Double Mask Representation*, April 2019, working paper or preprint, https://hal.inria.fr/hal-02102225

References in notes

[71] C. C. AGGARWAL, C. ZHAI (editors). *Mining Text Data*, Springer, 2012

[72] F. BAADER, D. CALVANESE, D. MCGUINNESS, D. NARDI, P. PATEL-SCHNEIDER (editors). *The Description Logic Handbook*, Cambridge University Press, Cambridge, UK, 2003

[73] M. ALAM, A. BUZMAKOV, V. CODOCEDO, A. NAPOLI. *Mining Definitions from RDF Annotations Using Formal Concept Analysis*, in "International Joint Conference in Artificial Intelligence", Buenos Aires, Argentina, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, July 2015, https://hal.archives-ouvertes.fr/hal-01186204

[74] M. ALAM, A. BUZMAKOV, A. NAPOLI. *Exploratory Knowledge Discovery over Web of Data*, in "Discrete Applied Mathematics", 2017, pp. 1-25, forthcoming, https://hal.inria.fr/hal-01673439

[75] M. ALAM, T. N. N. LE, A. NAPOLI. *LatViz: A New Practical Tool for Performing Interactive Exploration over Concept Lattices*, in "CLA 2016 - Thirteenth International Conference on Concept Lattices and Their Applications", Moscow, Russia, July 2016, https://hal.inria.fr/hal-01420751

[76] M. ATENCIA, J. DAVID, J. EUZENAT. *Data interlinking through robust linkkey extraction*, in "Proceedings of the 21st European Conference on Artificial Intelligence (ECAI)", T. SCHAUB, G. FRIEDRICH, B. O'SULLIVAN (editors), IOS Press, 2014, pp. 15–20, ftp://ftp.inrialpes.fr/pub/exmo/publications/atencia2014b.pdf

[77] J. BAIXERIES, V. CODOCEDO, M. KAYTOUE, A. NAPOLI. *Characterizing Approximate-Matching Dependencies in Formal Concept Analysis with Pattern Structures*, in "Discrete Applied Mathematics", 2018, vol. 249, pp. 18-27 [*DOI :* 10.1016/J.DAM.2018.03.073], https://hal.inria.fr/hal-01673441

[78] J. CANO, P. GUTIÉRREZ, B. KRAWCZYK, M. WOZNIAK, S. GARCÍA. *Monotonic classification: An overview on algorithms, performance measures and data sets*, in "Neurocomputing", 2019, vol. 341, pp. 168–182

[79] S. DA SILVA, F. LE BER, C. LAVIGNE. *Structures de haies dans un paysage agricole : une étude par chemin de Hilbert adaptatif et chaînes de Markov* , in "EGC 2016 – 16èemes Journées Francophones "Extraction et Gestion des Connaissances"", Reims, France, Revue des Nouvelles Technologies de l'Information, January 2016, vol. RNTI-E-30, pp. 279–290, https://hal.archives-ouvertes.fr/hal-01266344

[80] B. GANTER, S. O. KUZNETSOV. *Pattern Structures and Their Projections*, in "Proceedings of ICCS 2001", LNCS 2120, Springer, 2001, pp. 129–142

[81] M. GRABISCH, J.-L. MARICHAL, R. MESIAR, E. PAP. *Aggregation Functions*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2009

[82] O. HUDRY, B. MONJARDET. *Consensus Theories. An oriented survey*, in "Mathématiques et Sciences Humaines", 2010, vol. 190, n^{O} 2, pp. 139–167

[83] N. JUNIARTA, M. COUCEIRO, A. NAPOLI, C. RAÏSSI. *Sequence Mining within Formal Concept Analysis for Analyzing Visitor Trajectories*, in "SMAP 2018 - 13th International Workshop on Semantic and Social Media Adaptation and Personalization", Zaragoza, Spain, September 2018, https://hal.inria.fr/hal-01887927

[84] N. JUNIARTA, M. COUCEIRO, A. NAPOLI, C. RAÏSSI. *Sequential Pattern Mining using FCA and Pattern Structures for Analyzing Visitor Trajectories in a Museum*, in "CLA 2018 - The 14th International Conference on Concept Lattices and Their Applications", Olomouc, Czech Republic, June 2018, https://hal.inria.fr/hal-01887914

[85] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI, S. DUPLESSIS. *Mining gene expression data with pattern structures in formal concept analysis*, in "Information Sciences", 2011, vol. 181, n^{o} 10, pp. 1989-2001, https://hal.archives-ouvertes.fr/hal-00541100

[86] J.-F. MARI, A. GOBILLOT, M. BENOÎT. *Time Space Simulation of Land Use changes by stochastic modeling*, in "Revue Internationale de Géomatique", August 2018, vol. 28, n^{o} 2, pp. 219–242, https://hal.inria.fr/hal-01662140

[87] J.-P. METIVIER, A. LEPAILLEUR, A. BUZMAKOV, G. POEZEVARA, B. CRÉMILLEUX, S. O. KUZNETSOV, J. LE GOFF, A. NAPOLI, R. BUREAU, B. CUISSART. *Discovering structural alerts for mutagenicity using stable emerging molecular patterns*, in "Journal of Chemical Information and Modeling", 2015, vol. 55, n^{o} 5, pp. 925–940 [*DOI :* 10.1021/CI500611V], https://hal.archives-ouvertes.fr/hal-01186716

[88] G. PERSONENI, M.-D. DEVIGNES, M. SMAÏL-TABBONE, P. JONVEAUX, C. BONNET, A. COULET. *Cooperation of bio-ontologies for the classification of genetic intellectual disabilities : a diseasome approach*, in "Proceedings of the 11th International Conference on Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS 2018)", Antwerp, Belgium, December 2018, https://hal.inria.fr/hal-01925471

[89] N. RAMAKRISHNAN, D. KUMAR, B. MISHRA, M. POTTS, R. F. HELM. *Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions*, in "Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", New York, NY, USA, KDD '04, ACM, 2004, pp. 266–275

[90] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *Relational Concept Analysis: Mining Concept Lattices From Multi-Relational Data*, in "Annals of Mathematics and Artificial Intelligence", January 2013, vol. 67, n^{o} 1, pp. 81-108 [*DOI :* 10.1007/S10472-012-9329-3], http://hal.inria.fr/lirmm-00816300