



IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2016

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

1. Members	2
2. Overall Objectives	3
3. Research Program	3
3.1. Knowledge Discovery guided by Domain Knowledge	3
3.2. Text Mining	4
3.3. Knowledge Systems and Web of Data	4
4. Application Domains	5
4.1. Life Sciences: Biology, Chemistry and Medicine	5
4.2. Cooking	6
4.3. Agronomy	6
4.4. Digital Humanities	6
5. Highlights of the Year	7
6. New Software and Platforms	7
6.1. Symbolic KDD Systems	7
6.1.1. LatViz	7
6.1.2. OrphaMine – Data mining platform for orphan diseases	7
6.1.3. POQEMON Analytics: Platform for Quality Evaluation of Mobile Networks	8
6.1.4. Siren - Interactive and visual redescription mining	8
6.2. Stochastic systems for knowledge discovery and simulation	8
7. New Results	9
7.1. The Mining of Complex Data	9
7.1.1. FCA and Variations: RCA, Pattern Structures and Biclustering	9
7.1.2. Text Mining	9
7.1.3. Mining Sequences and Trajectories	10
7.1.4. Redescription Mining	10
7.1.5. E-sports analytics and subgroup discovery based on a single-player game	10
7.1.6. Data Privacy: Online link disclosure strategies for social networks	11
7.1.7. Aggregation	11
7.2. Knowledge Discovery in Healthcare and Life Sciences	12
7.2.1. Ontology-based Clustering of Biological Linked Open Data	12
7.2.2. Biological Data Aggregation for Knowledge Discovery	12
7.2.3. Suggesting Valid Pharmacogenes by Mining Linked Open Data and Electronic Health Records	13
7.2.4. Analysis of biomedical data annotated with ontologies	13
7.3. Knowledge Engineering and Web of Data	13
7.3.1. Around the Taaable Research Project	13
7.3.2. Exploring and Classifying the Web of Data	14
7.4. Advances in Graph Theory	14
8. Partnerships and Cooperations	15
8.1. Regional Initiatives	15
8.2. National Initiatives	15
8.2.1. ANR	15
8.2.1.1. Hybride (2011-2016)	15
8.2.1.2. ISTEEX (2014–2016)	16
8.2.1.3. PractiKPharma (2016–2020)	16
8.2.1.4. Termith (2014–2016)	16
8.2.2. FUI POQEMON (2014-2016)	16
8.2.3. CNRS PEPS and Mastodons projects	17
8.2.3.1. Mastodons HyQual (2016–2018)	17

8.2.3.2.	PEPS Confocal (2015–2016)	17
8.2.3.3.	PEPS Prefute (2015–2016)	17
8.3.	European Initiatives	17
8.4.	International Initiatives	18
8.4.1.	Inria Associate Teams:SNOWFLAKE	18
8.4.2.	Participation in Other International Programs	18
8.4.2.1.	A stay at NASA Frontier Development Lab	18
8.4.2.2.	Ciência Sem Fronteiras (2014–2016)	19
8.4.2.3.	STIC AmSud: Autonomic Knowledge Discovery (AKD, 2015–2016)	19
8.4.2.4.	LEA STRUCO	19
8.4.2.5.	Research Collaboration with HSE Moscow	19
9.	Dissemination	20
9.1.	Promoting Scientific Activities	20
9.2.	Teaching – Supervision – Juries	20
10.	Bibliography	20

Project-Team ORPAILLEUR

Creation of the Project-Team: 2008 January 01

Keywords:

Computer Science and Digital Science:

- 3. - Data and knowledge
 - 3.1.1. - Modeling, representation
 - 3.1.7. - Open data
 - 3.2. - Knowledge
 - 3.2.1. - Knowledge bases
 - 3.2.2. - Knowledge extraction, cleaning
 - 3.2.3. - Inference
 - 3.2.4. - Semantic Web
 - 3.2.5. - Ontologies
 - 3.3.2. - Data mining
 - 3.3.3. - Big data analysis
 - 3.4.1. - Supervised learning
 - 3.4.2. - Unsupervised learning
 - 3.4.5. - Bayesian methods
 - 3.4.8. - Deep learning
 - 3.5.2. - Recommendation systems
- 4. - Security and privacy
 - 4.1. - Threat analysis
 - 7.2. - Discrete mathematics, combinatorics
 - 7.9. - Graph theory
 - 8. - Artificial intelligence
 - 8.1. - Knowledge
 - 8.2. - Machine learning
 - 8.6. - Decision support

Other Research Topics and Application Domains:

- 1.1.2. - Molecular biology
- 1.2. - Ecology
 - 1.2.1. - Biodiversity
- 2. - Health
- 2.3. - Epidemiology
 - 2.4.1. - Pharmacokinetics and dynamics
 - 2.4.2. - Drug resistance
- 3.1. - Sustainable development
- 3.5. - Agronomy
- 3.6. - Ecology
 - 3.6.1. - Biodiversity
- 6.3.4. - Social Networks

- 6.4. - Internet of things
- 8.5.2. - Crowd sourcing
- 9. - Society and Knowledge
- 9.4.5. - Data science

1. Members

Research Scientists

Amedeo Napoli [Team leader, CNRS, Senior Researcher, HDR]
Esther Galbrun [Researcher, Inria]
Chedy Raïssi [Researcher, Inria]
Jean-Sébastien Sereni [Researcher, CNRS]

Faculty Members

Miguel Couceiro [Univ. Lorraine, Professor, HDR]
Adrien Coulet [Univ. Lorraine, Associate Professor]
Nicolas Jay [Univ. Lorraine, Professor, HDR]
Florence Le Ber [ENGEES Strasbourg, Professor (Team Associate), HDR]
Jean Lieber [Univ. Lorraine, Associate Professor, HDR]
Jean-François Mari [Univ. Lorraine, Professor, HDR]
Emmanuel Nauer [Univ. Lorraine, Associate Professor]
Sébastien Da Silva [Univ. Lorraine, Associate Professor]
Malika Smaïl-Tabbone [Univ. Lorraine, Associate Professor, HDR]
Yannick Toussaint [Univ. Lorraine, Professor (since Sep 2016, formerly Inria Researcher), HDR]

Engineers

Jérémie Nevin [Univ. Lorraine]
Thi Nhu Nguyen Le [Inria]
Mickaël Zehren [Inria, until May 2016]

PhD Students

Quentin Brabant [Univ. Lorraine]
Kévin Dalleau [Univ. Lorraine, since Dec 2016]
Rémi de Joannis de Verclos [Univ. Grenoble-Alpes]
Nyoman Juniarta [Univ. Lorraine, since Dec 2016]
Pierre Monnin [Univ. Lorraine, since Sep 2016]
François Pirot [Univ. Lorraine/Univ. Nijmegen, since Oct 2016]
Justine Reynaud [Inria]
Mohsen Sayed [Univ. Lorraine]
My Thao Tang [Univ. Lorraine, until June 2016]

Post-Doctoral Fellows

Emmanuelle Gaillard [Univ. Lorraine, ATER]
Dhouha Grissa [INRA]
Joël Legrand [Univ. Lorraine, since Nov 2016]

Visiting Scientist

Lucien Haddad [Visiting Scientist, until Mar 2016]

Administrative Assistants

Antoinette Courier [CNRS]
Emmanuelle Deschamps [Inria]
Sylvie Musilli [Univ. Lorraine]

Others

Aurore Alcolei [ENS Lyon, Student, until March 2016]
Marie Blemus [Inria, Student, from Jun 2016 until Aug 2016]
Yelen Per [CNRS, Student (Alternance)]

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.

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 related to knowledge bases (or domain ontologies) related to the domain of data for implementing *knowledge discovery guided by domain knowledge* or KDDK. In the KDDK process, the extracted units have “a life” after the interpretation step: 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, reifying the complementarity of knowledge discovery and knowledge representation.

3. Research Program

3.1. Knowledge Discovery guided by Domain Knowledge

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining formal concept analysis, classification, pattern mining second-order Hidden Markov Models

Knowledge discovery in databases (KDD) is aimed at discovering patterns in large databases. These patterns can then be interpreted as knowledge units to be reused in knowledge 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. 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 [80]) and extensions of FCA such as Pattern Structures [83] and Relational Concept Analysis (RCA [90]). Numerical methods are based on probabilistic approaches such as second-order Hidden Markov Models (HMM [85]), which are well adapted to the mining of temporal and spatial data.

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 operational sequence “mining, interpreting (modeling), representing, and reasoning”. In this way, knowledge discovery appears as a core task in knowledge engineering, with an impact in various semantic activities, e.g. information retrieval, recommendation and ontology engineering. Usual application domains for the team include agronomy, astronomy, biology, chemistry, and medicine.

One main operation in the research work of Orpailleur on KDDK is *classification*, which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. Classification problems can be formalized by means of a class of objects (or individuals), a class of attributes (or properties), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Formal Concept Analysis (FCA) relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting a set of formal concepts then organized within a concept lattice [80] (concept lattices are also known as “Galois lattices” [68]).

In parallel, the search for frequent itemsets and the extraction of association rules are well-known symbolic data mining methods, related to FCA (actually searching for frequent itemsets can be understood as traversing a concept lattice). Both processes usually produce a large number of items and rules, leading to the associated problems of “mining the sets of extracted items and rules”. Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow finding interesting subsets of association rules, e.g. informative association rules. This explains why several algorithms are needed for mining data depending on specific applications [92].

3.2. Text Mining

Keywords: text mining, knowledge discovery from collection of texts, annotation, ontology engineering from texts

The objective of a text mining process is to extract useful knowledge units from large collections of texts [78]. 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 particular task. To avoid information dispersion, 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 KDDK perspective, text mining is aimed at extracting “interesting units” (nouns and relations) from texts with the help of domain knowledge encoded within an ontology (also useful for text annotation). Text mining is especially useful in the context of semantic web for ontology engineering. In the Orpailleur team, the focus is put on the mining of real-world texts in application domains such as biology and medicine, using mainly symbolic data mining methods, and especially Formal Concept Analysis. 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

The web of data constitutes a good platform for experimenting ideas on knowledge engineering and knowledge discovery, in relation with the principles of semantic web. 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. The knowledge representation language recommended by W3C to design ontologies and knowledge bases is OWL, which is based on description logics (DLs [65]). In OWL, knowledge units are represented by classes (DL concepts) having properties (DL roles) and instances. Concepts are organized within a partial order based on a subsumption relation, and the inference services are based on classification-based reasoning and case-based reasoning (CBR).

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 be adapted for designing concept definitions in ontologies. Accordingly, we are interested here in two main challenges: how the web of data, as a set of potential knowledge sources (e.g. DBpedia, Wikipedia, Yago, Freebase...) can be mined for helping the design of definitions and knowledge bases and how knowledge discovery techniques can be applied for providing a better usage of the web of data (e.g. LOD classification).

Accordingly, 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 [66]. We have proposed a novel technique based on FCA which allows data exploration as well as the discovery of definition (bidirectional implication rules).

4. Application Domains

4.1. Life Sciences: Biology, Chemistry and Medicine

Participants: Adrien Coulet, Nicolas Jay, Joël Legrand, Jean Lieber, Pierre Monnin, Amedeo Napoli, Chedy Raïssi, Mohsen Sayed, Malika Smaïl-Tabbone, Yannick Toussaint, Mickaël Zehren.

Keywords: knowledge discovery in life sciences, bioinformatics, biology, chemistry, medicine, pharmacogenomics

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 environment, or between genetic and phenotypic data, especially for public health and pharmacogenomics domains. The latter case appears to be one main challenge in knowledge discovery in biology and involves knowledge discovery from complex data depending on domain knowledge.

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 among which 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 undirected labeled graphs. Graph mining methods may play an important role in this framework and Formal Concept Analysis can also be used in an efficient and well-founded way [86]. Graph mining in the framework of FCA is a very important task on which we are working, whose results can be transferred to text mining as well.

We are working on knowledge management in medicine and analysis of patient trajectories. The Kasimir research project is about decision support and knowledge management for the treatment of cancer. This is a multidisciplinary research project in which researchers in computer science (Orpailleur) and experts in oncology are participating. For a given cancer localization, a treatment is based on a protocol, which is applied in 70% of the cases and provides a treatment. The 30% remaining cases are “out of the protocol”, e.g. contraindication, treatment impossibility, etc. and the protocol should be adapted, based on discussions among specialists. This adaptation process is modeled in Kasimir thanks to CBR, where semantic web technologies are used and adapted for several years.

The analysis of patient trajectories, i.e. the “path” of a patient during illness (chronic illnesses and cancer), can be considered as an analysis of sequences. It is important to understand such sequential data and sequence mining methods should be adapted for addressing the complex nature of medical events. We are interested in the analysis of trajectories at different levels of granularity and w.r.t. external domain ontologies. In addition, it is also important to be able to compare and classify trajectories according to their content. Then we are also interested in the definition of similarity measures able to take into account the complex nature of trajectories and that can be efficiently implemented for allowing quick and reliable classifications.

PractiKPharma (Practice-based evidences for actioning Knowledge in Pharmacogenomics) is a starting research project about the validation of state-of-the-art knowledge in pharmacogenomics by mining “Electronic Health Records” (EHRs) [55]. Pharmacogenomics is a field studying how genomic variations impact drug responses. Most of the state of the art in the field is only available in biomedical literature, with various levels of validation. Accordingly we propose firstly, to extract pharmacogenomic knowledge units from the literature and secondly, to confirm or moderate these units by mining EHRs. Comparing knowledge units extracted from the literature with facts extracted from EHRs is not a trivial task for several reasons, among which (i) the literature is in English, whereas EHRs are in French, (ii) EHRs represent observations at the patient level whereas the literature is generalizing sets of patients...

4.2. Cooking

Participants: Emmanuelle Gaillard, Jean Lieber, Emmanuel Nauer.

Keywords: cooking, knowledge engineering, case-based reasoning, semantic web

123 The origin of the Taaable project is the Computer Cooking Contest (CCC). A contestant to CCC is a system that answers queries about recipes, using a recipe base; if no recipe exactly matches the query, then the system adapts another recipe. Taaable is a case-based reasoning system based on various technologies from semantic web, knowledge discovery, knowledge representation and reasoning. From a research viewpoint the system enables to test scientific results and to study the complementarity of various research trends in an application domain which is simple to understand and which raises complex issues at the same time.

4.3. Agronomy

Participants: Sébastien Da Silva, Florence Le Ber, Jean-François Mari.

Keywords: simulation in agronomy, graph model in agronomy

Research in agronomy was conducted in the framework of an Inria-INRA collaboration, taking place in the INRA research network PAYOTE about landscape modeling. In this framework, Sébastien da Silva prepared and defended a PhD thesis [74] in September 2014, supervised by Claire Lavigne (DR in ecology, INRA Avignon) and Florence Le Ber. The research work was related to the characterization and the simulation of hedgerow structures in agricultural landscapes, based on Hilbert-Peano curves and Markov models [48].

Moreover, an on-going research work about the representation of peasant knowledge is involved within a collaboration with IRD in Madagascar [81]. Sketches drawn by peasants were transformed into graphs and compared thanks to Formal Concept Analysis.

4.4. Digital Humanities

Participant: Jean Lieber.

Keywords: digital humanities, semantic web, SPARQL, approximate search, case-based reasoning

Recent contacts with the digital humanity community occurred with a group of researchers working in history and philosophy of science and technologies (located in Brest, Montpellier and Nancy). They want to benefit from semantic Web technologies in order to provide better accesses to their text corpora. A paper based on this starting collaboration was published [69], about exact and approximate search in RDFS-annotated text corpora based on the SPARQL technology and on case-based reasoning principles.

5. Highlights of the Year

5.1. Highlights of the Year

- The conference paper got the best paper award at the International Conference on Concept Lattices and Applications 2016 in Moscow, July 2016 (<https://cla2016.hse.ru/awards>). This reward was given to the paper and also to the whole work on the formalization of functional dependencies done by the four authors during the last years.
- In July 2016, Chedy Raïssi visited NASA Ames and SETI Institute as part of the Frontier Development Lab. He worked there for six weeks on the planetary defense community and focused on Delay-Doppler radar imaging. This stay was organized in the framework of the NASA “Asteroid Grand Challenge” program, where participation is based on a strong selection process.

BEST PAPER AWARD:

[33]

V. CODOCEDO, J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Characterization of Order-like Dependencies with Formal Concept Analysis*, in "Thirteenth International Conference on Concept Lattices and Their Applications (CLA 2016)", Moscou, Russia, M. HUCHARD, S. O. KUZNETSOV (editors), CEUR-WS.org, July 2016, vol. 1624, pp. 123-134, <https://hal.archives-ouvertes.fr/hal-01348496>

6. New Software and Platforms

6.1. Symbolic KDD Systems

6.1.1. LatViz

- Contact: Thi Nhu Nguyen Le
- URL: <http://latviz.loria.fr/latviz/>
- KEYWORDS: Formal Concept Analysis, Pattern Mining, Concept Lattice, Implications, Visualization

FUNCTIONAL DESCRIPTION.

LatViz is a new tool which allows the construction, the display and the exploration of concept lattices. LatViz proposes some remarkable improvements over existing tools and introduces various new functionalities focusing on interaction with experts, such as visualization of pattern structures (for dealing with complex non-binary data), AOC-posets (the core elements of the lattice), concept annotations, filtering based on various criteria and a visualization of implications [28], [27]. 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.

6.1.2. OrphaMine – Data mining platform for orphan diseases

- Partners: INSERM – MoDYCo CNRS – Greyc Université de Caen Basse Normandie
- Contact: Chedy Raïssi
- URL: <http://orphamine.inria.fr/>
- KEYWORDS: Bioinformatics, data mining, biology, health, data visualization, drug development.

FUNCTIONAL DESCRIPTION.

The OrphaMine platform, developed as part of the ANR Hybrid project, enables visualization, data integration and in-depth analytics. The data at the heart of the platform is about orphan diseases and is extracted from the OrphaData ontology (<http://www.orpha.net>).

We aim at building a true collaborative portal that will serve the different actors of the Hybrid project: (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.

6.1.3. *POQEMON Analytics: Platform for Quality Evaluation of Mobile Networks*

- Partners: Altran, DataPublica, GenyMobile, HEC, Inria Nancy-Grand Est, IP-Label, Next Interactive Media, Orange, Université Paris-Est Créteil
- Contact: Chedy Raïssi
- URL: <https://members.loria.fr/poqemon/>
- KEYWORDS: Data mining, data visualization.

FUNCTIONAL DESCRIPTION.

POQEMON is a quality evaluation platform for mobile phone networks developed in the Orpailleur team in the framework of an FUI project (see 8.2.2). The quality measures which are studied include the coverage, availability and network performances. Multiple methods are implemented in this platform, either in visualization or in data anonymization to make on-line analytics as simple as possible.

6.1.4. *Siren - Interactive and visual redescription mining*

- Contact: Esther 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 redescrptions. 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 redescrptions is not enough. The expert must also be able to understand the redescrptions found, adjust them to better match his domain knowledge and test alternative hypotheses with them, for instance. Thus, Siren allows mining redescrptions 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 redescrptions.

6.2. Stochastic systems for knowledge discovery and simulation

6.2.1. *The CarottAge and ARPEnTAge Systems*

- Contact: Jean-François Mari
- URL: <http://carottage.loria.fr>
- 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 [84]. CarottAge is currently used by INRA researchers interested in mining the changes in territory and landscape related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination. CarottAge was also used for mining hydromorphological data and gave interesting results for that purpose.

ARPEntAge, for “Analyse de Régularités dans les Paysages : Environnement, Territoires, Agronomie” is built on top of the CarottAge system to fully take into account the spatial dimension of input sequences. It can be used for analyzing spatio-temporal databases [85] and for space-time clustering of a landscape based on temporal land uses. Displaying tools and the generation of time-dominant shape files have also been defined. With agronomists, we are now focusing on the simulation of unknown spatial time sequences in order to explore various crop management scenarios.

CarottAge and ARPEntAge are freely available under GPL license. A special effort is currently aimed at designing interactive visualization tools to provide the expert a user-friendly interface.

7. New Results

7.1. The Mining of Complex Data

Participants: Quentin Brabant, Miguel Couceiro, Adrien Coulet, Esther Galbrun, Nicolas Jay, Nyoman Juniarta, Florence Le Ber, Joël Legrand, Pierre Monnin, Amedeo Napoli, Justine Reynaud, Chedy Raïssi, Mohsen Sayed, My Thao Tang, Yannick Toussaint.

Keywords: formal concept analysis, relational concept analysis, pattern structures, pattern mining, association rule, redescription mining, graph mining, sequence mining, biclustering, aggregation

Pattern mining and Formal Concept Analysis are suitable symbolic methods for KDDK, that may be used for real-sized applications. Global improvements are carried out on the scope of applicability, the ease of use, the efficiency of the methods, and on the ability to fit evolving situations. Accordingly, the team is extending these symbolic data mining methods for working on complex data (e.g. textual documents, biological, chemical or medical data), involving objects with multi-valued attributes (e.g. domains or intervals), n-ary relations, sequences, trees and graphs.

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 data. In particular, FCA, which usually applies to binary data-tables, can be adapted to work on more complex data. In this way, we have contributed to two main extensions of FCA, namely Pattern Structures and Relational Concept Analysis. Pattern Structures (PS [79]) allow building a concept lattice from complex data, e.g. numbers, sequences, trees and graphs. Relational Concept Analysis (RCA) is able to analyze objects described both by binary and relational attributes [90] 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 designed new information retrieval methods based on FCA [72], text classification and heterogeneous pattern structures [71], pattern structures for structured attribute sets [67], and also a quasi-polynomial algorithm for mining top patterns w.r.t. measures satisfying special properties in a FCA framework [70]. We developed also a whole line of work on pattern structures for the discovery of functional dependencies [33], text classification and heterogeneous pattern structures [71], and fuzzy FCA as well [31]. Finally, we also proposed new visualization techniques and tools able to display important and useful information (e.g. stable concepts) from large concept lattices [28].

7.1.2. Text Mining

The thesis work of My Thao Tang [11] proposes a process where software and humans agents cooperate in knowledge discovery from different source textual types for extending a knowledge base. One challenge is that, on the one hand, knowledge discovery methods (software agents) can be run in accordance with background knowledge (or expert knowledge), at any step of the KDD process. On the other hand, human agents should be able to correct or to extend the current knowledge base. FCA is used for discovering a “class schema” (or “representation model”) within textual resources which can be either a set of attribute implications or

a concept lattice. However, such a schema does not necessarily fit the point of view of a domain expert for different reasons, e.g. noise, errors or exceptions in the data. Thus, a bridge filling the possible gap between the representation model based on a concept lattice and the representation model of a domain expert is studied in [44]. The background knowledge is encoded as a set of attribute dependencies or constraints which is “aligned” with the set of implications associated with the concept lattice. Such an alignment may lead to modifications in the original concept lattice. This method can be generalized for generating lattices satisfying some constraints based on attribute dependencies in using the so-called “extensional projections”. It also allows experts to keep a trace of the changes occurring in the original lattice and the revised version, and to assess how concepts in practice are related to concepts discovered in the data.

In the framework of the Hybride ANR project (see 8.2.1.1), Mohsen Sayed proposes an original machine learning approach for identifying in literature disease phenotypes that are not yet represented within existing ontologies. The process is based on graph patterns extracted from sentences represented as dependency graphs. Phenotypes are usually expressed by complex noun phrases while traditional gazetteers recognize them only partially. The strength of graph patterns is to preserve the linguistic component bounds and to enable the identification of the complete phenotype formulation. A specific publication is currently in preparation.

7.1.3. Mining Sequences and Trajectories

Nowadays data sets are available in very complex and heterogeneous ways. Mining of such data collections is essential to support many real-world applications ranging from healthcare to marketing. This year, we completed a research work on the analysis of “complex sequential data” by means of interesting sequential patterns [13]. We approach the problem using FCA and pattern structures, where the subsumption relation ordering patterns is defined w.r.t. the partial order on sequences.

7.1.4. Redescription Mining

Among the mining methods developed in the team is redescription mining. 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 [89]. It is motivated by the idea that in scientific investigations data oftentimes have different nature. For instance, 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 that exist between these different aspects.

A practical example in biology consists in finding geographical areas that admit two characterizations, one in terms of their climatic profile and one in terms of the occupying species. Discovering such redescrptions can contribute to better our understanding of the influence of climate over species distribution. Besides biology, applications of redescription mining can be envisaged in medicine or sociology, among other fields.

In recent work [40], we focused on the problem of pattern selection, developing a method for filtering a set of redescription to identify a non-redundant, interesting subset to present to the analyst. Also, we showcased the usability of redescription mining on an application in the political domain [50]. More specifically, we applied redescription mining to the exploratory analysis of the profiles and opinions of candidates to the parliamentary elections in Finland in 2011 and 2015.

We presented an introductory tutorial on redescription mining at ECML-PKDD in September 2016 to help foster the research on these techniques and widen their use (<http://siren.mpi-inf.mpg.de/tutorial/main/>).

7.1.5. E-sports analytics and subgroup discovery based on a single-player game

Discovering patterns that strongly distinguish one class label from another is a challenging data-mining task. The unsupervised discovery of such patterns would enable the construction of intelligible classifiers and to elicit interesting hypotheses from the data. Subgroup Discovery (SD) [87] is one framework that formally defines this pattern mining task. However, SD still faces two major issues: (i) how to define appropriate quality measures to characterize the uniqueness of a pattern; (ii) how to select an accurate heuristic search technique when exhaustive enumeration of the pattern space is unfeasible. The first issue has been tackled by the Exceptional Model Mining (EMM) framework [77]. This general framework aims to find patterns that

cover tuples that locally induce a model that substantially differs from the model of the whole dataset. The second issue has been studied in SD and EMM mainly with the use of beam-search strategies and genetic algorithms for discovering a pattern set that is non-redundant, diverse and of high quality. In [58], we argue that the greedy nature of most of these approaches produce pattern sets that lack of diversity. Consequently, we proposed to formally define pattern mining as a single-player game, as in a puzzle, and to solve it with a Monte Carlo Tree Search (MCTS), a recent technique mainly used for artificial intelligence and planning problems. The exploitation/exploration trade-off and the power of random search of MCTS lead to an any-time mining approach, in which a solution is always available, and which tends towards an exhaustive search if given enough time and memory. Given a reasonable time and memory budget, MCTS quickly drives the search towards a diverse pattern set of high quality. MCTS does not need any knowledge of the pattern quality measure, and we show to what extent it is agnostic to the pattern language.

7.1.6. Data Privacy: Online link disclosure strategies for social networks

Online social networks are transforming our culture and world. While online social networks have become an important channel for social interactions, they also raise ethical and privacy issues. A well known fact is that social networks leak information, that may be sensitive, about users. However, performing accurate real world online privacy attacks in a reasonable time frame remains a challenging task. In [57], [26] (this work is done in cooperation with the Pesto Inria Team), we address the problem of rapidly disclosing many friendship links using only legitimate queries (i.e., queries and tools provided by the targeted social network). Our study sheds new light on the intrinsic relation between communities (usually represented as groups) and friendships between individuals. To develop an efficient attack we analyzed group distributions, densities and visibility parameters from a large sample of a social network. By effectively exploring the target group network, our proposed algorithm is able to perform friendship and mutual-friend attacks along a strategy that minimizes the number of queries. The results of attacks performed on a major social network profiles show that 5 different friendship links are disclosed in average for each single legitimate query in the best cases.

7.1.7. Aggregation

Aggregation theory is the study of 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 [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, as well as methods for extracting them. Typical approaches include rough-set and FCA approaches, that we aim to extend in order to increase expressivity, applicability and readability of results. Direct applications of these efforts are expected in the context of two multidisciplinary projects, namely the “Fight Heart Failure” and the European H2020 “CrossCult” project.

In our recent work, 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, we provided axiomatizations of noteworthy classes of lattice-based aggregation functions, which were then used to model preferences and to provide their logical description [14]. In this qualitative setting, we also tackled the problem of computing version spaces (with explicit descriptions of all models compatible with a given dataset) and proved a dichotomy theorem showing that the problem is NP-complete for preferences over at least 4 attributes whereas it is solvable in polynomial time otherwise [61].

Finding consensual structures among different classifications or metrics is again a challenging task, especially, for large and multi-source data, and its importance becomes apparent since algorithmic approaches are often heuristic on such datasets and they rarely produce the same output. The difficulty in extracting such consensual structures is then to find appropriate and meaningful aggregation rules, and their impossibility is often revealed by Arrow type impossibility results. In the current year, we focused on median structures [19], [21] that include several relational structures (trees, graphs, lattices) and allow several consensus procedures.

7.2. Knowledge Discovery in Healthcare and Life Sciences

Participants: Miguel Couceiro, Adrien Coulet, Kévin Dalleau, Joël Legrand, Pierre Monnin, Amedeo Napoli, Chedy Raïssi, Mohsen Sayed, Malika Smaïl-Tabbone, Yannick Toussaint.

Life Sciences constitute a challenging domain for KDDK. Biological data are complex from many points of views, e.g. voluminous, high-dimensional and deeply inter-connected. Analyzing such data is a crucial issue in healthcare, environment and agronomy. Besides, many bio-ontologies are available and can be used to enhance the knowledge discovery process. Accordingly, the research work of the Orpailleur team in KDDK applied to Life Sciences is in concern with mining biological data, data integration, information retrieval, and use of bio-ontologies and linked data for resource annotation.

7.2.1. Ontology-based Clustering of Biological Linked Open Data

Increasing amounts of biomedical data provided as Linked Open Data (LOD) offer novel opportunities for knowledge discovery in bio-medicine. We proposed an approach for selecting, integrating, and mining LOD with the goal of discovering genes responsible for a disease [88]. We are currently working on the integration of LOD about known phenotypes and genes responsible for diseases along with relevant bio-ontologies. We are also defining a corpus-based semantic distance. One possible application of this work is to build and compare possible diseaseomes, i.e. global graphs representing all diseases connected according to their pairwise similarity values.

7.2.2. Biological Data Aggregation for Knowledge Discovery

Two multi-disciplinary projects were initiated in 2016, in collaboration with the Capsid Team, with a group of clinicians from the Regional University Hospital (CHU Nancy) and bio-statisticians from the Maths Lab (IECL). The first project is entitled ITM2P ¹ and depends on the so-called CPER 2015–2020 framework. We are involved in the design of the SMEC platform as a support for “Simulation, Modeling and Knowledge Extraction from Bio-Medical Data”.

The second project is a RHU ² project entitled *Fight Heart Failure* (FHF), where we are in charge of a workpackage about “data aggregation” mechanisms. Accordingly, we are working on the definition of multidimensional similarity measure for comparing and clustering sets of patients. Each cluster should correspond to a bioprofile, i.e. a subgroup of patients sharing the same form of the disease and thus the same diagnosis and care strategy.

The first results were presented at the International Symposium on Aggregation and Structures (ISAS 2016) [36]. We propose the GABS for “Graph Aggregation Based Similarity” approach for complex graph aggregation resulting in a similarity graph between a subset of nodes. Indeed the initial graph contains two types of nodes, i.e. individuals and attributes. The pairwise similarity between individuals is derived from the various paths in the initial graph. This setting allows the integration of domain knowledge in the initial graph (corresponding to domain ontologies, norms...). Another advantage of the GABS approach is to generate a similarity graph which can be used as input for various clustering algorithms (graph-based ones as well as those working on similarity/distance matrix).

¹“Innovations Technologiques, Modélisation et Médecine Personnalisée”

²“Recherche Hospitalo-Universitaire”

The next question will be to build a prediction model for each bioprofile/subgroup (once validated by the clinicians) for a decision support system. Thus, we are investigating SRL (“Statistical Relational Learning”) methods which combine symbolic and probabilistic methods for improving expressivity (through logical or relational languages) and for dealing with uncertainty.

7.2.3. *Suggesting Valid Pharmacogenes by Mining Linked Open Data and Electronic Health Records*

A standard task in pharmacogenomics research is identifying genes that may be involved in drug response variability and called “pharmacogenes”. As genomic experiments in this domain tend to generate many false positives, computational approaches based on background knowledge may generate more valuable results. Until now, the later have only used molecular networks databases or biomedical literature. We are studying a new method taking advantage of various linked data sources to validate uncertain drug-gene relationships, i.e. pharmacogenes [75]. One advantage relies on the standard implementation of linked data that facilitates the joint use of various sources and makes easier the consideration of features of various origins. Accordingly, we selected, formatted, interconnected and published an initial set of linked data sources relevant to pharmacogenomics. We applied numerical classification methods for extracting drug-gene pairs that can become validated pharmacogene candidates.

The ANR project “PractiKPharma” initiated in 2016 relies on similar ideas, having the motivation of validating state-of-the-art knowledge in pharmacogenomics (<http://praktikpharma.loria.fr/>). The originality of “PractiKPharma” is to use Electronic Health Records (EHRs) to constitute cohorts of patients that can be mined for validating extracted pharmacogenomics knowledge units.

7.2.4. *Analysis of biomedical data annotated with ontologies*

In the context of the Snowflake Inria Associate team, Gabin Personeni, who is a PhD student co-supervised by Marie-Dominique Devignes (Capsid EPI) and Adrien Coulet (Orpailleur EPI) spent four months at the Stanford University in 2016. After this internship, we developed an approach based on pattern structures to identify frequently associated ADRs (Adverse Drug Reactions) from patient data either in the form of EHR or ADR spontaneous reports [51], [49]. In this case, pattern structures provide an expressive representation of ADR, taking into account the multiplicity of drugs and phenotypes involved in such reactions. Additionally, pattern structures allow considering diverse biomedical ontologies used to represent or annotate patient data, enabling a “semantic” comparison of ADRs. Up to now, this is the first research work considering such representations to mine rules between frequently associated ADRs. We illustrated the generality of the approach on two distinct patient datasets, each of them linked to distinct biomedical ontologies. The first dataset corresponds to anonymized EHRs, extracted from “STRIDE”, the EHR data warehouse of Stanford Hospital and Clinics. The second dataset is extracted from the U.S. FDA (for Food & Drug Administration) Adverse Event Reporting System (FAERS). Several significant association rules have been extracted and analyzed and may be used as a basis of a recommendation system.

7.3. Knowledge Engineering and Web of Data

Participants: Emmanuelle Gaillard, Nicolas Jay, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer, Justine Reynaud.

Keywords: knowledge engineering, web of data, definition mining, classification-based reasoning, case-based reasoning, belief revision, semantic web

7.3.1. *Around the Taaable Research Project*

The Taaable project was originally created as a challenger of the Computer Cooking Contest (ICCB Conference) [73]. Beyond its participation to the CCC challenges, the Taaable project aims at federating various research themes: case-based reasoning (CBR), information retrieval, knowledge acquisition and extraction, knowledge representation, belief change theory, ontology engineering, semantic wikis, text-mining, etc. CBR performs adaptation of recipes w.r.t. user constraints. The reasoning process is based on a cooking domain ontology (especially hierarchies of classes) and adaptation rules. The knowledge base is encoded within a semantic wiki containing the recipes, the domain ontology and adaptation rules.

As acquiring knowledge from experts is costly, a new approach was proposed to allow a CBR system using partially reliable, non expert, knowledge from the web for reasoning. This approach is based on notions such as belief, trust, reputation and quality, as well as their relationships and rules to manage the knowledge reliability. The reliability estimation is used to filter knowledge with high reliability as well as to rank the results produced by the CBR system. Performing CBR with knowledge resulting from an e-community is improved by taking into account the knowledge reliability [10]. In the same way, another study shows how the case retrieval of a CBR system can be improved using typicality. Typicality discriminates subclasses of a class in the domain ontology depending of how a subclass is a good example for its class. An approach has been proposed to partition the subclasses of some classes into atypical, normal and typical subclasses in order to refine the domain ontology. The refined ontology allows a finer-grained generalization of the query during the retrieval process, improving at the same time the final results of the CBR system.

The Taaable system also includes a module for adapting textual preparations (from a source recipe text to an adapted recipe text, through a formal representation in the qualitative algebra INDU). The evaluation of this module as a whole thanks to users has been carried out and has shown its efficiency (w.r.t. text quality and recipe quality), when compared with another approach to textual adaptation [76].

FCA allows the classification of objects according to the properties they share into a concept lattice. A lattice has been built on a large set a cooking recipes according to the ingredients they use, producing a hierarchy of ingredient combinations. When a recipe R has to be adapted, this lattice can be used to search the best ingredient combinations in the concepts that are the closest to the concept representing R .

Minimal change theory and belief revision can be used as tools to support adaptation in CBR, i.e. the source case is modified to be consistent with the target problem using a revision operator. Belief revision was applied to Taaable to adjust the ingredient quantities using specific inference engines.

Another approach to adaptation based on the principles of analogical transfer applied to the formalism RDFS has been developed [41]. It is based on the problem-solution dependency represented as an RDFS graph: this dependency within the source case is modified so that it fits the context of the target problem. The application problem that has guided this research addresses the issue of cocktail name adaptation: given a cocktail recipe, the name of this cocktail and the ingredient substitution that produces a new cocktail, how could the new cocktail be called?

7.3.2. *Exploring and Classifying the Web of Data*

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. The popularization and 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. Particularly, we are interested in the completeness of the data and the their potential to provide concept definitions in terms of necessary and sufficient conditions [66]. We have proposed a novel technique based on Formal Concept Analysis which organizes subsets of RDF data into a concept lattice [43]. This allows data exploration as well as the discovery of implication rules which are used to automatically detect missing information and then to complete RDF data and to provide definitions. Moreover, this is also a way of reconciling syntax and semantics in the LOD cloud. Experiments on the DBpedia knowledge base shows that this kind of approach is well-founded and effective.

7.4. **Advances in Graph Theory**

Participants: Aurore Alcolei, Rémi de Joannis de Verclos, François Pirot, Jean-Sébastien Sereni.

Keywords: graph theory, graph coloring, extremal graph theory, chromatic number, two-mode data networks

Motivated by notions brought forward by sociology, we confirm a conjecture by Everett, Sinclair, and Dankelmann [Some Centrality results new and old, *J. Math. Sociology* 28 (2004), 215–227] regarding the problem of maximizing closeness centralization in two-mode data, where the number of data of each type is fixed. Intuitively, our result states that among all networks obtainable via two-mode data, the largest closeness is achieved by simply locally maximizing the closeness of a node. Mathematically, our study concerns bipartite graphs with fixed size bipartitions, and we show that the extremal configuration is a rooted tree of depth 2, where neighbors of the root have an equal or almost equal number of children [24].

Using recently introduced techniques based on entropy compression, we proved that the acyclic chromatic number of a graph with maximum degree Δ is less than $2.835\Delta^{4/3} + \Delta$. This improved the previous upper bound, which was $50\Delta^{4/3}$ (see [91] which is now published in *Journal of Combinatorics*, 7(4):725–737, 2016).

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. *Hydreos*

Participant: Jean-François Mari.

Hydreos is a state organization –actually a so-called “Pôle de compétitivité”– aimed at evaluating the delivering and the quality of water (<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 (see § 6.2.1) are used in this context, especially by agronomists of INRA (ASTER Mirecourt <http://www6.nancy.inra.fr/sad-aster>).

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. *Hybride (2011-2016)*

Participants: Adrien Coulet, Amedeo Napoli, Chedy Raïssi, My Thao Tang, Mohsen Sayed, Yannick Toussaint.

The Hybride research project (<http://hybride.loria.fr/>) aims at combining Natural Language Processing (NLP) and Knowledge Discovery in Databases (KDD) for text mining. A key idea is to design an interacting and convergent process where NLP methods are used for guiding text mining while KDD methods are used for guiding the analysis of textual documents. NLP methods are mainly based on text analysis and extraction of general and temporal information. KDD methods are based on pattern mining, e.g. patterns and sequences, formal concept analysis and graph mining. In this way, NLP methods applied to texts extract “textual information” that can be used by KDD methods as constraints for focusing the mining of textual data. By contrast, KDD methods extract patterns and sequences to be used for guiding information extraction from texts and text analysis. Experimental and validation parts associated with the Hybride project are provided by an application to the documentation of rare diseases in the context of Orphanet.

The partners of the Hybride consortium are the GREYC Caen laboratory (pattern mining, NLP, text mining), the MoDyCo Paris laboratory (NLP, linguistics), the INSERM Paris laboratory (Orphanet, ontology design), and the Orpailleur team at Inria NGE (FCA, knowledge representation, pattern mining, text mining). The Hybride project ended on 30th November 2016.

8.2.1.2. ISTEEX (2014–2016)

Participant: Yannick Toussaint.

ISTEX is a so-called “Initiative d’excellence” managed by CNRS and DIST (“Direction de l’Information Scientifique et Technique”). ISTEEX aims at giving to the research and teaching community an on-line access to scientific publications in all the domains (<http://www.istex.fr/istex-excellence-initiative-of-scientific-and-technical-information/>). Thus ISTEEX requires a massive acquisition of documents such as journals, proceedings, corpus, databases...ISTEX-R is one research project within ISTEEX in which the Orpailleur team is involved, with two other partners, namely ATILF laboratory and INIST Institute (both located in Nancy). ISTEEX-R aims at developing new tools for querying full-text documentation, analyzing content and extracting information. A platform is under development to provide robust NLP tools for text processing, as well as methods in text mining and domain conceptualization.

8.2.1.3. PractiKPharma (2016–2020)

Participants: Adrien Coulet, Joël Legrand, Pierre Monnin, Amedeo Napoli, Malika Smaïl-Tabbone, Yannick Toussaint.

The ANR project PractiKPharma (<http://praktikpharma.loria.fr/>) is interested in the validation of domain knowledge in pharmacogenomics. The originality of PractiKPharma is to use “Electronic Health Records” (EHRs) to constitute cohorts of patients, cohorts which are then mined for validating extracted pharmacogenomics knowledge units after validation w.r.t. literature knowledge. This project involves two other labs, namely LIRMM at Montpellier and CRC Paris.

8.2.1.4. Termith (2014–2016)

Participant: Yannick Toussaint.

Termith (<http://www.atilf.fr/ressources/termith/>) is an ANR Project involving a series of laboratories, namely ATILF, INIST, Inria Nancy Grand Est, Inria Saclay, LIDILEM, and LINA. It aims at indexing documents belonging to different domain of Humanities. Thus, the project focuses on extracting candidate terms (information extraction) and on disambiguation.

In the Orpailleur team, we are mainly concerned by information extraction using Formal Concept Analysis techniques, but also pattern and sequence mining. The objective is to define contexts introducing terms, i.e. finding textual environments allowing a system to decide whether a textual element is actually a candidate term and its corresponding environment. This disambiguation process was described and published at LREC 2016 [35]. The Termith project ended in April 2016.

8.2.2. FUI POQEMON (2014-2016)

Participants: Chedy Raïssi, Mickaël Zehren.

The publication of transaction data, such as market basket data, medical records, and query logs, serves the public benefit. Mining such data allows the derivation of association rules that connect certain items to others with measurable confidence. Still, this type of data analysis poses a privacy threat; an adversary having partial information on a person’s behavior may confidently associate that person to an item deemed to be sensitive. Ideally, an anonymization of such data should lead to an inference-proof version that prevents the association of individuals to sensitive items, while otherwise allowing truthful associations to be derived. The POQEMON project aims at developing new pattern mining methods and tools for supporting privacy preserving knowledge discovery from monitoring purposes on mobile phone networks. The main idea is to develop sound approaches that handle the tradeoff between privacy of data and the power of analysis. Original approaches to this problem were based on value perturbation, damaging data integrity. Recently, value generalization has been proposed as an alternative; still, approaches based on it have assumed either that all items are equally sensitive, or that some are sensitive and can be known to an adversary only by association, while others are non-sensitive and can be known directly. Yet in reality there is a distinction between sensitive and non-sensitive items, but an adversary may possess information on any of them. Most critically, no antecedent method aims at a clear inference-proof privacy guarantee. In this project, we integrated the ρ -uncertainty privacy concept that inherently safeguards against sensitive associations without constraining the nature of an adversary’s

knowledge and without falsifying data. The project integrates the ρ -uncertainty pattern mining approach with novel data visualization techniques.

The POQEMON research project (<https://members.loria.fr/poqemon/>) involves the following partners: Altran, DataPublica, GenyMobile, HEC, IP-Label, Next Interactive Media, Orange and Université Paris-Est Créteil, and Inria Nancy Grand Est.

8.2.3. CNRS PEPS and Mastodons projects

8.2.3.1. Mastodons HyQual (2016–2018)

Participants: Miguel Couceiro, Esther Galbrun, Dhouha Grissa, Amedeo Napoli, Chedy Raïssi, Justine Reynaud.

The HyQual project was proposed and initiated this year in the framework of the Mastodons CNRS Call about data quality in data mining (see <http://www.cnrs.fr/mi/spip.php?article819&lang=fr>). This project is interested in the mining of nutritional data for discovering predictive biomarkers of diabetes and metabolic syndrome in elder populations. The data mining methods which are considered here are hybrid, combining symbolic and numerical methods, and are applied to complex and noisy metabolic data [39]. In the HyQual project, we are mainly interested by the quality of the data at hand and the patterns that can be discovered. In particular, we check whether we can find possible definitions within the data (actually double implications) and redescrptions (under the form of different descriptions of the same data). In this way, we can study the definitional power of the data and as well the incompleteness of the data, leading to two original ways of considering data quality. The project involves researchers from the Orpailleur Team, with researchers from LIRIS Lyon, ICube Strasbourg, and INRA Clermont-Ferrand.

8.2.3.2. PEPS Confocal (2015–2016)

Participants: Adrien Coulet, Amedeo Napoli, Chedy Raïssi, Malika Smaïl-Tabbone.

The Confocal Project (see <http://www.cnrs.fr/ins2i/spip.php?article1183>) is interested in the design of new methods in bioinformatics for analyzing and classifying heterogeneous omics data w.r.t. biological domain knowledge. We are working on the adaption of FCA and pattern structures for discovering patterns and associations in gene data with the help of domain ontologies. One important objective of the project is to check whether such a line of research could be reused on so-called “discrete models in molecular biology”.

8.2.3.3. PEPS Prefute (2015–2016)

Participants: Quentin Brabant, Adrien Coulet, Miguel Couceiro, Esther Galbrun, Amedeo Napoli, Chedy Raïssi, Justine Reynaud, Mohsen Sayed, Malika Smaïl-Tabbone, My Thao Tang, Yannick Toussaint.

The PEPS Prefute project is mainly interested in interaction and iteration in the knowledge discovery (KD) process. Usually the KD process is organized around three main steps which are (i) selection and preparation of the data, (ii) data mining, and (iii) interpretation of (selected) resulting patterns. An analyst, most of the time an expert of the data domain, is present for leading the KD process. Accordingly, the PEPS Prefute project is interested in the study of interactions between the analyst and the KD process, i.e. pushing constraints, preferences and domain knowledge, for guiding and improving the KD process. One possible way is to discover initial patterns acting as seeds for searching farther the pattern space w.r.t. this initial seeds possibly linked to preferences of the analyst. In this way, the interesting pattern space is much more concise and of much lower size.

Then, the importance of preferences and domain knowledge in interaction with KD, and as well, visualization tools, have to be improved for allowing work with large and complex datasets (see <https://www.greyc.fr/fr/node/2207>).

8.3. European Initiatives

8.3.1. FP7 & H2020 Projects

8.3.1.1. CrossCult (H2020 Project, 2016-2020)

Participants: Miguel Couceiro, Nyoman Juniarta, Amedeo Napoli, Chedy Raïssi.

CrossCult (<http://www.crosscult.eu/>) aims to make 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. The first goal is to lower cultural EU barriers and create unique cross-border perspectives, by connecting existing digital historical resources and by creating new ones through the participation of the public. The second goal is to provide long-lasting experiences of social learning and entertainment that will help towards the better understanding and re-interpretation of European history. To achieve these goals, CrossCult will use cutting-edge technology to connect existing digital cultural assets and to combine them with interactive experiences that all together are intended to increase retention, stimulate reflection and help European citizens appreciate their past and present in a holistic manner. CrossCult will be implemented on 4 real-world flagship pilots involving a total of 8 sites across Europe.

The role of the Orpailleur Team (in conjunction with the Kiwi Team of LORIA) is mainly to work on the recommendation aspects, with a focus on defining an extended profile of the users and connecting these profiles with domain knowledge for leading the recommendation process [42].

The partners of the Orpailleur team in the CrossCult project are the following: Luxembourg Institute for Science and Technology and Centre Virtuel de la Connaissance sur l'Europe (Luxembourg, leader 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 Guías Interactivas (Spain).

8.4. International Initiatives

8.4.1. Inria Associate Teams: SNOWFLAKE

8.4.1.1. SNOWFLAKE

Participants: Adrien Coulet, Joël Legrand, Pierre Monnin, Malika Smaïl-Tabbone.

Title: Knowledge Discovery from Linked Data and Clinical Notes

International Partner (Institution - Laboratory - Researcher):

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

Start year: 2014

Web site: <http://snowflake.loria.fr/>

Snowflake is an Inria Associate Team which started in 2014. It is aimed at facilitating the collaboration between researchers from the Inria Orpailleur team and the Stanford Center for Biomedical Informatics Research, Stanford University, USA. The main objective of Snowflake is to improve biomedical knowledge discovery by connecting Electronic Health Records (EHRs) with domain knowledge either in the form of ontologies or of Linked Open Data (LOD). Such a connection should help to complete domain knowledge w.r.t. EHRs. The initial focus of Snowflake is the identification and characterization of groups of patients w.r.t. (adverse) reactions to drugs. Identified features associated with such groups of patients could be used as predictors of over- or under-reactions to some drugs.

8.4.2. Participation in Other International Programs

8.4.2.1. A stay at NASA Frontier Development Lab

Participant: Chedy Raïssi.

In 2013, NASA presented the "Asteroid Grand Challenge", a White House supported initiative to supplement the NEO (Near-Earth Object) Program, with a mission: "Find all asteroid threats to human populations and to know what to do about them." There remain a number of unresolved gaps in this challenge, both in terms of discovery, characterization and eventual mitigation strategies, should a potentially hazardous asteroid (PHA) be discovered. By bringing new approaches in computer science, such as deep learning and data mining to tackle specific parts of the problem, solutions may be revealed that, combined with existing processes, significantly benefit the community as a whole.

In July 2016, Chedy Raïssi visited NASA Ames and SETI Institute as part of the Frontier Development Lab. He worked there on developing meaningful research opportunities, as well as support the work of the planetary defense community and show the potential of this kind of applied research methodology to deliver breakthrough of significant value. The work was over a period of six weeks, focusing on Delay-Doppler radar imaging. Delay-Doppler radar imaging is a powerful technique to characterize the trajectories, shapes, and spin states of near-Earth asteroids and has yielded detailed models of dozens of objects. Since the 1990s, Delay-Doppler data has been analyzed using the SHAPE software developed originally by Steven J. Ostro. SHAPE performs sequential single-parameter fitting, and requires considerable computation runtime and human intervention. Reconstructing asteroid shapes and spins from Delay-Doppler data is, like many inverse problems, computationally intensive and requires extensive human oversight of the shape modeling process. Chedy Raïssi has explored two new techniques to better automate Delay-Doppler shape modeling: Bayesian optimization and deep generative models.

8.4.2.2. *Ciência Sem Fronteiras (2014–2016)*

Participant: Amedeo Napoli.

Program “Ciência Sem Fronteiras” is a Brazilian research fellowship which provides a funding for the stay of a visiting French researcher in Brazil at Universidade Federal Pernambuco Recife for three years. The on-going project is called “Formal Concept Analysis as a Support for Knowledge Discovery” and is aimed at combining FCA methods with numerical clustering methods used by Brazilian colleagues. This project is supervised in Brazil by Professor Francisco de A.T. de Carvalho (CIn/UFPE).

The project aims at developing and comparing classification and clustering algorithms for complex data (especially interval and multi-valued data). Two families of algorithms are studied, namely “clustering algorithms” based on the use of a similarity or a distance for comparing the objects, and “classification algorithms in Formal Concept Analysis (FCA)” based on attribute sharing between objects. The objectives here are to combine the facilities of both families of algorithms for improving the potential of each family in dealing with more complex and voluminous datasets.

8.4.2.3. *STIC AmSud: Autonomic Knowledge Discovery (AKD, 2015–2016)*

Participants: Miguel Couceiro, Esther Galbrun, Amedeo Napoli, Chedy Raïssi.

This research project involves researchers with different specialties, from Brazil (Universidade Federal Rio Grande do Sul), from Chile (UFESM Santiago and Valparaiso), from Uruguay (Universidad de la República), and the Orpailleur Team. The project is interested in the design of solutions able to proactively understand the behavior of systems and networks in order to prevent vulnerable states. Accordingly, we aim at integrating knowledge discovery techniques within autonomic systems in order to provide intelligent self-configuration and self-protection mechanisms. The results of this project may not only benefit to end-users but also highly contribute to the scientific community by providing solid foundations for the development of more secure, scalable, and reliable management approaches.

8.4.2.4. *LEA STRUCO*

Participant: Jean-Sébastien Sereni.

LEA STRUCO is an “Associated International Laboratory” of CNRS between IÚUK, Prague, and LIAFA, Paris. It focuses on high-level study of fundamental combinatorial objects, with a particular emphasis on comprehending and disseminating the state-of-the-art theories and techniques developed. The obtained insights shall be applied to obtain new results on existing problems as well as to identify directions and questions for future work. Jean-Sébastien Sereni is the contact person for LEA STRUCO which was initiated when Jean-Sébastien was a member of LIAFA.

8.4.2.5. *Research Collaboration with HSE Moscow*

Participants: Miguel Couceiro, Adrien Coulet, Amedeo Napoli, Chedy Raïssi, Justine Reynaud.

An on-going 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 visited Inria Nancy Grand Est several times too. The collaboration is materialized by the joint supervision of students (such as for example the thesis of Aleksey Buzmakov defended at the end of 2015), and the organization of scientific events, as in particular the workshop FCA4AI whose fifth edition was organized this year in August at ECAI 2016 (see <http://www.fca4ai.hse.ru>) [53]. A special session about Knowledge Discovery and Formal Concept Analysis will be supervised by Sergei O. Kuznetsov and Amedeo Napoli at the next ISMIS Conference in Warsaw (Poland) next June 2017 (http://ismis2017.ii.pw.edu.pl/s_kd_fca.php).

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific Events Organization, General Chairs, Scientific Chairs

- “FCA4AI 2016”. Amedeo Napoli organized with Sergei O. Kuznetsov (HSE Moscow) and Sebastian Rudolph (TU Dresden) the fifth workshop FCA4AI (“What can do FCA for Artificial Intelligence”) which was associated with the ECAI Conference in The Hague (Netherlands, August 2016, see <http://www.fca4ai.hse.ru/2016> and <http://ceur-ws.org/Vol-1703>) [53].

9.1.1.1. 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/?p=1>). 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 (ECAI, ECML-PKDD, ICCBR, ICDM, ICFCA, IJCAI, KDD...), as members of editorial boards, and finally in the organization of journal special issues.

9.2. Teaching – Supervision – Juries

- The members of the Orpailleur team are involved in teaching at all levels of teaching, mainly at University of Lorraine. Actually, most of the members of the Orpailleur team are employed on “Université de Lorraine” positions.
- The members of the Orpailleur team are also involved in student supervision, at all university levels, from under-graduate until post-graduate students.
- Finally, the members of the Orpailleur team are involved in HDR and thesis defenses, being thesis referees or thesis committee members.

10. Bibliography

Major publications by the team in recent years

- [1] E. BRESSO, R. GRISONI, G. MARCHETTI, A.-S. KARABOGA, M. SOUCHET, M.-D. DEVIGNES, M. SMAÏL-TABBONE. *Integrative relational machine-learning for understanding drug side-effect profiles*, in "BMC Bioinformatics", 2013, vol. 14, n^o 1, 207 p. [DOI : 10.1186/1471-2105-14-207], <https://hal.inria.fr/hal-00843914>

- [2] M. COUCEIRO, D. DUBOIS, H. PRADE, T. WALDHAUSER. *Decision-making with Sugeno integrals: Bridging the gap between multicriteria evaluation and decision under uncertainty*, in "Order", 2016, vol. 33, n^o 3, pp. 517-535 [DOI : 10.1007/s11083-015-9382-8], <https://hal.archives-ouvertes.fr/hal-01184340>
- [3] K. DALLEAU, N. COUMBA NDIAYE, A. COULET. *Suggesting valid pharmacogenes by mining linked data*, in "Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2015", Cambridge, United Kingdom, Proceedings of the Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2015, December 2015, <https://hal.inria.fr/hal-01239568>
- [4] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER, L. MARTIN. *Case Adaptation with Qualitative Algebras*, in "International Joint Conferences on Artificial Intelligence", Pékin, China, F. ROSSI (editor), Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, August 2013, pp. 3002-3006, <https://hal.inria.fr/hal-00871703>
- [5] Z. DVOŘÁK, J.-S. SERENI, J. VOLEC. *Subcubic triangle-free graphs have fractional chromatic number at most 14/5*, in "Journal of the London Mathematical Society", June 2014, vol. 89, n^o 3, pp. 641–662, <https://hal.archives-ouvertes.fr/hal-00779634>
- [6] E. EGHO, C. RAÏSSI, T. CALDERS, N. JAY, A. NAPOLI. *On measuring similarity for sequences of itemsets*, in "Data Mining and Knowledge Discovery", May 2015, vol. 29, n^o 3, 33 p. [DOI : 10.1007/s10618-014-0362-1], <https://hal.inria.fr/hal-01094383>
- [7] E. GALBRUN, A. GIONIS, N. TATTI. *Top-k overlapping densest subgraphs*, in "Data Mining and Knowledge Discovery", September 2016, vol. 30, n^o 5, pp. 1134 - 1165 [DOI : 10.1007/s10618-016-0464-z], <https://hal.archives-ouvertes.fr/hal-01399184>
- [8] M. KAYTOUE, V. CODOCEDO, A. BUZMAKOV, J. BAIXERIES, S. O. KUZNETSOV, A. NAPOLI. *Pattern Structures and Concept Lattices for Data Mining and Knowledge Processing*, in "Machine Learning and Knowledge Discovery in Databases", Porto, Portugal, A. BIFET, M. MAY, B. ZADROZNY, R. GAVALDA, D. PEDRESCHI, F. BONCHI, J. CARDOSO, M. SPILIOPOULOU (editors), Lecture Notes in Computer Science, Springer International Publishing, 2015, vol. 9286, pp. 227-231 [DOI : 10.1007/978-3-319-23461-8_19], <https://hal.archives-ouvertes.fr/hal-01188637>
- [9] J.-F. MARI, E.-G. LAZRAK, M. BENOÎT. *Time space stochastic modelling of agricultural landscapes for environmental issues*, in "Environmental Modelling and Software", March 2013, vol. 46, pp. 219-227 [DOI : 10.1016/j.envsoft.2013.03.014], <https://hal.inria.fr/hal-00807178>

Publications of the year

Doctoral Dissertations and Habilitation Theses

- [10] E. GAILLARD. *Managing and exploiting knowledge produced by an e-community – Application to case-based reasoning*, Université de Lorraine, CNRS, Inria, LORIA, UMR 7503, June 2016, <https://hal.inria.fr/tel-01415288>
- [11] M. T. TANG. *An Interactive and Iterative Knowledge Extraction Process Using Formal Concept Analysis*, Université de lorraine, June 2016, <https://hal.inria.fr/tel-01425357>

Articles in International Peer-Reviewed Journals

- [12] J. ALMEIDA, M. COUCEIRO, T. WALDHAUSER. *On the topological semigroup of equational classes of finite functions under composition*, in "Journal of Multiple-Valued Logic and Soft Computing", 2016, 24 p. , <https://hal.archives-ouvertes.fr/hal-01090645>
- [13] A. BUZMAKOV, E. EGHO, N. JAY, S. O. KUZNETSOV, A. NAPOLI, C. RAÏSSI. *On Mining Complex Sequential Data by Means of FCA and Pattern Structures*, in "Int. J. Gen. Syst.", 2016, vol. 45, n^o 2, pp. 135-159 [DOI : 10.1080/03081079.2015.1072925], <https://hal.archives-ouvertes.fr/hal-01186715>
- [14] M. COUCEIRO, D. DUBOIS, H. PRADE, T. WALDHAUSER. *Decision-making with sugeno integrals: Bridging the gap between multicriteria evaluation and decision under uncertainty*, in "Order", 2016, vol. 33, n^o 3, pp. 517-535 [DOI : 10.1007/s11083-015-9382-8], <https://hal.archives-ouvertes.fr/hal-01184340>
- [15] M. COUCEIRO, L. HADDAD, I. G. ROSENBERG. *Partial clones containing all Boolean monotone self-dual partial functions*, in "Journal of Multiple-Valued Logic and Soft Computing", 2016, vol. 27, n^o 2-3, pp. 183-192, <https://hal.archives-ouvertes.fr/hal-01093942>
- [16] M. COUCEIRO, L. HADDAD, K. SCHÖLZEL, T. WALDHAUSER. *A Solution to a Problem of D. Lau: Complete Classification of Intervals in the Lattice of Partial Boolean Clones*, in "Journal of Multiple-Valued Logic and Soft Computing", 2016, 8 p. , <https://hal.inria.fr/hal-01183004>
- [17] M. COUCEIRO, L. HADDAD, K. SCHÖLZEL, T. WALDHAUSER. *On the interval of strong partial clones of Boolean functions containing $Pol((0,0), (0,1), (1,0))$* , in "Algebra Universalis", 2016, <https://hal.inria.fr/hal-01404588>
- [18] M. COUCEIRO, E. LEHTONEN. *On the arity gap of finite functions : results and applications.*, in "Journal of Multiple-Valued Logic and Soft Computing", 2016, vol. 27, n^o 1-2, 15 p. , <https://hal.inria.fr/hal-01175695>
- [19] M. COUCEIRO, J.-L. MARICHAL, B. TEHEUX. *Conservative Median Algebras and Semilattices*, in "Order", March 2016, vol. 33, n^o 1, pp. 121-132 [DOI : 10.1007/s11083-015-9356-x], <https://hal.inria.fr/hal-01175696>
- [20] M. COUCEIRO, J.-L. MARICHAL, B. TEHEUX. *Relaxations of associativity and preassociativity for variadic functions*, in "Fuzzy Sets and Systems", 2016, vol. 229, pp. 89-97, <https://hal.archives-ouvertes.fr/hal-01184334>
- [21] M. COUCEIRO, G. MELETIOU. *On a special class of median algebras*, in "Miskolc Mathematical Notes", 2016, <https://hal.inria.fr/hal-01248142>
- [22] E. GALBRUN, A. GIONIS, N. TATTI. *Top-k overlapping densest subgraphs*, in "Data Mining and Knowledge Discovery", September 2016, vol. 30, n^o 5, pp. 1134 - 1165 [DOI : 10.1007/s10618-016-0464-z], <https://hal.archives-ouvertes.fr/hal-01399184>
- [23] D. GRISSA, M. PÉTÉRA, M. BRANDOLINI, A. NAPOLI, B. COMTE, E. PUJOS-GUILLOT. *Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data*, in "Frontiers in Molecular Biosciences", June 2016, vol. 3, 15 p. [DOI : 10.3389/FMOLB.2016.00030], <https://hal.archives-ouvertes.fr/hal-01421002>
- [24] M. KRNC, J.-S. SERENI, R. SKREKOVSKI, Z. YILMA. *Closeness Centralization Measure for Two-mode Data of Prescribed Sizes*, in "Network Science", 2016, <https://hal.inria.fr/hal-01289470>

- [25] J.-S. SERENI, J. VOLEC. *A note on acyclic vertex-colorings*, in "Journal of Combinatorics", September 2016, vol. 7, n^o 4, pp. 725–737, <https://hal.archives-ouvertes.fr/hal-00921122>

International Conferences with Proceedings

- [26] Y. ABID, A. IMINE, A. NAPOLI, C. RAÏSSI, M. RUSINOWITCH. *Online link disclosure strategies for social networks*, in "The 11th International Conference on Risks and Security of Internet and Systems", Roscoff, France, The 11th International Conference on Risks and Security of Internet and Systems, September 2016, <https://hal.inria.fr/hal-01402062>
- [27] M. ALAM, T. N. N. LE, A. NAPOLI. *Interactive Exploration over Concept Lattices with LatViz*, in "European Conference on Artificial Intelligence, Software Demo Track.", The Hague, Netherlands, August 2016, <https://hal.inria.fr/hal-01420752>
- [28] 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>
- [29] M. ALAM, T. N. N. LE, A. NAPOLI. *Steps Towards Interactive Formal Concept Analysis with LatViz*, in "Proceedings of the 5th International Workshop "What can FCA do for Artificial Intelligence"? co-located with the European Conference on Artificial Intelligence ECAI 2016", The Hague, Netherlands, August 2016, <https://hal.inria.fr/hal-01420753>
- [30] A. BUZMAKOV, A. NAPOLI. *How Fuzzy FCA and Pattern Structures are connected?*, in "5th Workshop "What can FCA do for Artificial Intelligence?" (FCA4AI'2016)", The Hague, Netherlands, August 2016, <https://hal.inria.fr/hal-01420997>
- [31] A. BUZMAKOV, A. NAPOLI. *On Scaling of Fuzzy FCA to Pattern Structures*, in "The 13th International Conference on Concept Lattices and their Applications (CLA2016)", Moscow, Russia, July 2016, <https://hal.inria.fr/hal-01421000>
- [32] C. CHAMARD-JOVENIN, A. CHESNEL, E. BRESSO, C. MOREL, C. THIÉBAUT, M. SMAÏL-TABBONE, E.-H. DJERMOUNE, M.-D. DEVIGNES, T. BOUKHOBZA, H. DUMOND. *Transgenerational effects of ER-alpha36 over-expression on mammary gland development and molecular phenotype: clinical perspective for breast cancer risk and therapy.*, in "21st World Congress on Advances in Oncology and 19th International Symposium on Molecular Medicine", Athens, Greece, October 2016, <https://hal.archives-ouvertes.fr/hal-01416469>
- [33] *Best Paper*
V. CODOCEDO, J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Characterization of Order-like Dependencies with Formal Concept Analysis*, in "Thirteenth International Conference on Concept Lattices and Their Applications (CLA 2016)", Moscou, Russia, M. HUCHARD, S. O. KUZNETSOV (editors), CEUR-WS.org, July 2016, vol. 1624, pp. 123-134, <https://hal.archives-ouvertes.fr/hal-01348496>.
- [34] V. CODOCEDO, J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Contributions to the Formalization of Order-like Dependencies using FCA*, in "What can FCA do for Artificial Intelligence?", The Hague, Netherlands, August 2016, <https://hal.inria.fr/hal-01420630>

- [35] B. DAILLE, E. JACQUEY, G. LEJEUNE, L. F. MELO, Y. TOUSSAINT. *Ambiguity Diagnosis for Terms in Digital Humanities*, in "Language Resources and Evaluation Conference", Portorož, Slovenia, May 2016, <https://hal.inria.fr/hal-01423650>
- [36] K. DALLEAU, M. COUCEIRO, M.-D. DEVIGNES, C. RAÏSSI, M. SMAÏL-TABBONE. *Using aggregation functions on structured data: a use case in the FIGHT-HF project*, in "International Symposium on Aggregation and Structures (ISAS 2016)", Luxembourg, Luxembourg, G. KISS, J.-L. MARICHAL, B. TEHEUX (editors), International Symposium on Aggregation and Structures (ISAS 2016) - Book of abstracts, July 2016, <https://hal.inria.fr/hal-01399232>
- [37] D. GRISSA, B. COMTE, E. PUJOS-GUILLOT, A. NAPOLI. *A Hybrid Approach for Mining Metabolomic Data*, in "FCA4AI@ECAI 2016", La Haye, Netherlands, Proceedings of the 5th International Workshop "What can FCA do for Artificial Intelligence"? co-located with the European Conference on Artificial Intelligence, FCA4AI@ECAI 2016, CEUR-WS.org, August 2016, vol. 1703, <https://hal.archives-ouvertes.fr/hal-01422050>
- [38] D. GRISSA, B. COMTE, E. PUJOS-GUILLOT, A. NAPOLI. *A Hybrid Data Mining Approach for the Identification of Biomarkers in Metabolomic Data*, in "Concept Lattices and Their Applications", Moscou, Russia, Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016, CEUR-WS.org, July 2016, vol. 1624, <https://hal.archives-ouvertes.fr/hal-01421015>
- [39] D. GRISSA, B. COMTE, E. PUJOS-GUILLOT, A. NAPOLI. *A Hybrid Knowledge Discovery Approach for Mining Predictive Biomarkers in Metabolomic Data*, in "ECML PKDD", Riva del garda, Italy, September 2016, pp. 572 - 587 [DOI : 10.1007/978-3-319-46128-1_36], <https://hal.archives-ouvertes.fr/hal-01421011>
- [40] J. KALOFOLIAS, E. GALBRUN, P. MIETTINEN. *From Sets of Good Redescriptions to Good Sets of Redescriptions*, in "Proceedings of the 16th IEEE International Conference on Data Mining, ICDM'16", Barcelona, Spain, December 2016, <https://hal.archives-ouvertes.fr/hal-01399294>
- [41] N. KIANI, J. LIEBER, E. NAUER, J. SCHNEIDER. *Analogical Transfer in RDFS, Application to Cocktail Name Adaptation*, in "International Conference on Case-Based Reasoning (ICCBR-2016)", Atlanta, United States, A. GOEL, M. B. DIAZ-AGUDO, T. ROTH-BERGHOFER (editors), Case-Based Reasoning Research and Development, 24th International Conference, ICCBR 2016, Springer, October 2016, vol. 9969, pp. 218 - 233 [DOI : 10.1007/978-3-319-47096-2_15], <https://hal.inria.fr/hal-01410237>
- [42] P.-E. OSCHÉ, S. CASTAGNOS, A. NAPOLI, Y. NAUDET. *Walk the line: Toward an efficient user model for recommendations in museums*, in "11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2016)", Thessaloniki, Greece, October 2016, pp. 83 - 88 [DOI : 10.1109/SMAP.2016.7753389], <https://hal.inria.fr/hal-01421548>
- [43] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Contribution to the Classification of Web of Data based on Formal Concept Analysis*, in "What can FCA do for Artificial Intelligence (FCA4AI) (ECAI 2016)", La Haye, Netherlands, August 2016, <https://hal.inria.fr/hal-01420466>
- [44] M. T. TANG, A. BUZMAKOV, Y. TOUSSAINT, A. NAPOLI. *Building a Domain Knowledge Model Based on a Concept Lattice Integrating Expert Constraints*, in "Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications", Moscow, Russia, July 2016, vol. 1624, pp. 349–362, <https://hal.inria.fr/hal-01400452>

National Conferences with Proceedings

- [45] Q. BRABANT, M. COUCEIRO. *Apprentissage d'intégrales de Sugeno à partir de données inconsistantes*, in "25èmes Rencontre Francophone sur la Logique Floue et ses Applications", La Rochelle, France, 25èmes Rencontres Francophones sur la Logique Floue et ses Applications, November 2016, pp. 49-56, <https://hal.archives-ouvertes.fr/hal-01404567>
- [46] Q. BRABANT, M. COUCEIRO, F. LABERNIA, A. NAPOLI. *Une approche de réduction de dimensionnalité pour l'agrégation de préférences qualitatives*, in "16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016", Reims, France, January 2016, pp. 345-350, <https://hal.inria.fr/hal-01404913>
- [47] C. CHAMARD-JOVENIN, A. CHESNEL, C. MOREL, M.-D. DEVIGNES, M. SMAÏL-TABBONE, T. BOUKHOBZA, H. DUMOND. *Long chain alkylphenol mixture promotes breast cancer initiation and progression through an ER α 36-mediated mechanism*, in "2nd French Workshop on Endocrine disruption in wildlife and human health", Paris, France, January 2016, Présentation Poster, <https://hal.archives-ouvertes.fr/hal-01320688>
- [48] 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 – 16èmes 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>
- [49] G. PERSONENI, M.-D. DEVIGNES, M. DUMONTIER, M. SMAÏL-TABBONE, A. COULET. *Extraction d'association d'EIM à partir de dossiers patients : expérimentation avec les structures de patrons et les ontologies*, in "Deuxième Atelier sur l'Intelligence Artificielle et la Santé", Montpellier, France, Atelier IA & Santé, June 2016, <https://hal.inria.fr/hal-01391172>

Conferences without Proceedings

- [50] E. GALBRUN, P. MIETTINEN. *Analysing Political Opinions Using Redescription Mining*, in "Proceedings of the Data Mining in Politics workshop at ICDM 2016, DMiP'16", Barcelona, Spain, December 2016, <https://hal.archives-ouvertes.fr/hal-01399254>
- [51] G. PERSONENI, M.-D. DEVIGNES, M. DUMONTIER, M. SMAÏL-TABBONE, A. COULET. *Discovering ADE associations from EHRs using pattern structures and ontologies*, in "Phenotype Day, Bio-Ontologies SIG, ISMB", Orlando, United States, July 2016, <https://hal.inria.fr/hal-01369448>

Scientific Books (or Scientific Book chapters)

- [52] A. W. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Classification and Exploration of 3D Protein Domain Interactions Using Kbdock*, in "Data Mining Techniques for the Life Sciences", O. CARUGO, F. EISENHABER (editors), Methods in Molecular Biology, Springer Science+Business Media New York, 2016, vol. 1415, pp. 91-105 [DOI : 10.1007/978-1-4939-3572-7_5], <https://hal.inria.fr/hal-01317448>

Books or Proceedings Editing

- [53] S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors). *Proceedings of the 5th International Workshop "What can FCA do for Artificial Intelligence?"*, FCA4AI 2016(co-located with ECAI 2016, The Hague, Netherlands,

August 30th 2016), CEUR Workshop Proceedings, CEUR-WS.org, The Hague, Netherlands, 2016, vol. 1703, 135 p. , <https://hal.inria.fr/hal-01425672>

Research Reports

- [54] K. EDWARDS, A. GIRÃO, J. VAN DEN HEUVEL, R. J. KANG, G. J. PULEO, J.-S. SERENI. *Extension from Precoloured Sets of Edges*, Loria & Inria Grand Est, 2016, <https://hal.archives-ouvertes.fr/hal-01024843>

Scientific Popularization

- [55] A. COULET, M. SMAÏL-TABBONE. *Mining Electronic Health Records to Validate Knowledge in Pharmacogenomics*, in "ERCIM News", January 2016, n^o 104, 56 p. , <https://hal.inria.fr/hal-01369501>
- [56] M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Kbdock - Searching and organising the structural space of protein-protein interactions*, in "ERCIM News", January 2016, n^o 104, pp. 24-25, <https://hal.inria.fr/hal-01258117>

Other Publications

- [57] Y. ABID, A. IMINE, A. NAPOLI, C. RAÏSSI, M. RIGOLOT, M. RUSINOWITCH. *Analyse d'activité et exposition de la vie privée sur les médias sociaux*, January 2016, 16ème conférence francophone sur l'Extraction et la Gestion des Connaissances (EGC 2016), Poster, <https://hal.inria.fr/hal-01241619>
- [58] G. BOSC, J.-F. BOULICAUT, C. RAÏSSI, M. KAYTOUE. *Anytime Discovery of a Diverse Set of Patterns with Monte Carlo Tree Search*, December 2016, working paper or preprint, <https://hal.inria.fr/hal-01418663>
- [59] M. COUCEIRO, M. GRABISCH. *On integer-valued means and the symmetric maximum*, September 2016, Documents de travail du Centre d'Economie de la Sorbonne 2016.80 - ISSN : 1955-611X, <https://halshs.archives-ouvertes.fr/halshs-01412025>
- [60] M. COUCEIRO, M. GRABISCH. *On integer-valued means and the symmetric maximum*, 2016, working paper or preprint, <https://hal.inria.fr/hal-01404593>
- [61] M. COUCEIRO, M. MARÓTI, T. WALDHAUSER, L. ZADORI. *Computing version spaces in the qualitative approach to multicriteria decision aid* , 2016, working paper or preprint, <https://hal.inria.fr/hal-01404590>
- [62] D. GRISSA, B. COMTE, E. PUJOS-GUILLOT, A. NAPOLI. *Combinaison de méthodes numériques et symboliques pour l'analyse de données métabolomiques*, 16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, 18-22 Janvier 2016, Reims, France, Hermann-Editions, January 2016, vol. E-30, pp. 555–556, Extraction et Gestion des Connaissances, Poster, <https://hal.archives-ouvertes.fr/hal-01421035>
- [63] J.-F. MARI, A. NAPOLI. *Maîtriser la complexité des systèmes: apport de la fouille de données*, November 2016, Cet exposé présente les travaux du projet Orpailleur en matière de fouille de données en agriculture numérique, <https://hal.inria.fr/hal-01414117>
- [64] J. REYNAUD, Y. TOUSSAINT, A. NAPOLI. *Analyse de concepts formels pour la classification de triplets RDF*, November 2016, Gestion de Données — Principes, Technologies et Applications (BDA 2016), Poster, <https://hal.archives-ouvertes.fr/hal-01420737>

References in notes

- [65] F. BAADER, D. CALVANESE, D. MCGUINNESS, D. NARDI, P. PATEL-SCHNEIDER (editors). *The Description Logic Handbook*, Cambridge University Press, Cambridge, UK, 2003
- [66] 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>
- [67] M. ALAM, A. BUZMAKOV, A. NAPOLI, A. SAILANBAYEV. *Revisiting Pattern Structures for Structured Attribute Sets*, in "Proceedings of the Twelfth International Conference on Concept Lattices and Their Applications Clermont-Ferrand, France, October 13-16, 2015.", Clermont-Ferrand, France, CEUR Workshop Proceedings, August 2015, vol. 1466, pp. 241-252, <https://hal.inria.fr/hal-01186339>
- [68] M. BARBUT, B. MONJARDET. *Ordre et classification – Algèbre et combinatoire (2 tomes)*, Hachette, Paris, 1970
- [69] O. BRUNEAU, S. GARLATTI, M. GUEDJ, S. LAUBÉ, J. LIEBER. *SemanticHPST: Applying Semantic Web Principles and Technologies to the History and Philosophy of Science and Technology*, in "The Semantic Web: ESWC 2015 Satellite Events", Portoroz, Slovenia, F. GANDON, A. ZIMMERMANN, C. FARON-ZUCKER, J. BRESLIN, S. VILLATA, C. GUÉRET (editors), Lecture Notes in Computer Science, Springer International Publishing, May 2015, vol. 9341, pp. 416-427 [DOI : 10.1007/978-3-319-25639-9_53], <https://hal.archives-ouvertes.fr/hal-01214698>
- [70] A. BUZMAKOV, S. O. KUZNETSOV, A. NAPOLI. *Fast Generation of Best Interval Patterns for Nonmonotonic Constraints*, in "Machine Learning and Knowledge Discovery in Databases", Porto, Portugal, Lecture Notes in Computer Science, September 2015, vol. 9285, pp. 157-172 [DOI : 10.1007/978-3-319-23525-7_10], <https://hal.archives-ouvertes.fr/hal-01186718>
- [71] V. CODOCEDO, A. NAPOLI. *A Proposition for Combining Pattern Structures and Relational Concept Analysis*, in "Formal Concept Analysis - 12th International Conference, ICFCA 2014, Cluj-Napoca, Romania, June 10-13, 2014. Proceedings", Cluj-Napoca, Romania, June 2014, pp. 96 - 111 [DOI : 10.1007/978-3-319-07248-7_8], <https://hal.inria.fr/hal-01095870>
- [72] V. CODOCEDO, A. NAPOLI. *Formal Concept Analysis and Information Retrieval – A Survey*, in "International Conference in Formal Concept Analysis - ICFCA 2015", Nerja, Spain, J. BAIXERIES, C. SACAREA, M. OJEDA-ACIEGO (editors), Formal Concept Analysis - Lecture Notes in Computer Science, Springer, June 2015, vol. 9113, pp. 61-77 [DOI : 10.1007/978-3-319-19545-2_4], <https://hal.archives-ouvertes.fr/hal-01186196>
- [73] A. CORDIER, V. DUFOUR-LUSSIER, J. LIEBER, E. NAUER, F. BADRA, J. COJAN, E. GAILLARD, L. INFANTE-BLANCO, P. MOLLI, A. NAPOLI, H. SKAF-MOLLI. *Taaable: a Case-Based System for personalized Cooking*, in "Successful Case-based Reasoning Applications-2", S. MONTANI, L. C. JAIN (editors), Studies in Computational Intelligence, Springer, January 2014, vol. 494, pp. 121-162 [DOI : 10.1007/978-3-642-38736-4_7], <https://hal.inria.fr/hal-00912767>
- [74] S. DA SILVA. *Spatial data mining and modelling of hedgrows in agricultural landscapes*, Université de Lorraine, September 2014, <https://hal.inria.fr/tel-01101424>

- [75] K. DALLEAU, N. COUMBA NDIAYE, A. COULET. *Suggesting valid pharmacogenes by mining linked data*, in "Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2015", Cambridge, United Kingdom, Proceedings of the Semantic Web Applications and Tools for Life Sciences (SWAT4LS) 2015, December 2015, <https://hal.inria.fr/hal-01239568>
- [76] V. DUFOUR-LUSSIER, J. LIEBER. *Evaluating a textual adaptation system*, in "International Conference on Case-Based Reasoning", Frankfurt, Germany, September 2015, <https://hal.inria.fr/hal-01178331>
- [77] W. DUIVESTIJN, A. FEELDERS, A. J. KNOBBE. *Exceptional Model Mining - Supervised descriptive local pattern mining with complex target concepts*, in "Data Mining and Knowledge Discovery", 2016, vol. 30, n^o 1, pp. 47–98, <http://dx.doi.org/10.1007/s10618-015-0403-4>
- [78] R. FELDMAN, J. SANGER. *The Text Mining Handbook (Advanced Approaches in Analyzing Unstructured Data)*, Cambridge University Press, 2007
- [79] B. GANTER, S. O. KUZNETSOV. *Pattern Structures and Their Projections*, in "Proceedings of ICCS 2001", LNCS 2120, Springer, 2001, pp. 129–142
- [80] B. GANTER, R. WILLE. *Formal Concept Analysis*, Springer, Berlin, 1999
- [81] D. HERVÉ, J.-H. RAMAROSON, A. RANDRIANARISON, F. LE BER. *Comment les paysans du corridor forestier de Fianarantsoa (Madagascar) dessinent-ils leur territoire ? Des cartes individuelles pour confronter les points de vue*, in "Cybergeographie : Revue européenne de géographie / European journal of geography", 2014, 681 p. , <https://hal.archives-ouvertes.fr/hal-01057112>
- [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] M. KAYTOUE, V. CODOCEDO, A. BUZMAKOV, J. BAIXERIES, S. O. KUZNETSOV, A. NAPOLI. *Pattern Structures and Concept Lattices for Data Mining and Knowledge Processing*, in "Machine Learning and Knowledge Discovery in Databases", Porto, Portugal, A. BIFET, M. MAY, B. ZADROZNY, R. GAVALDA, D. PEDRESCHI, F. BONCHI, J. CARDOSO, M. SPILIOPOULOU (editors), Lecture Notes in Computer Science, Springer International Publishing, 2015, vol. 9286, pp. 227-231 [DOI : 10.1007/978-3-319-23461-8_19], <https://hal.archives-ouvertes.fr/hal-01188637>
- [84] J.-F. MARI. *CarottAge Windows pour les données Teruti : manuel d'utilisation*, Loria & Inria Grand Est, February 2014, 43 p. , <https://hal.inria.fr/hal-00951102>
- [85] J.-F. MARI, F. LE BER, E.-G. LAZRAK, M. BENOÎT, C. ENG, A. THIBESSARD, P. LEBLOND. *Using Markov Models to Mine Temporal and Spatial Data*, in "New Fundamental Technologies in Data Mining", K. FUNATSU, K. HASEGAWA (editors), Intech, 2011, pp. 561–584, <http://hal.inria.fr/inria-00566801/en>
- [86] 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>

- [87] P. K. NOVAK, N. LAVRAC, G. I. WEBB. *Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining*, in "Journal of Machine Learning Research", 2009, vol. 10, pp. 377–403, <http://doi.acm.org/10.1145/1577069.1577083>
- [88] G. PERSONENI, S. DAGET, C. BONNET, P. JONVEAUX, M.-D. DEVIGNES, M. SMAÏL-TABBONE, A. COULET. *Mining Linked Open Data: A Case Study with Genes Responsible for Intellectual Disability*, in "Data Integration in the Life Sciences - 10th International Conference, DILS 2014", Lisbon, Portugal, H. GALHARDAS, E. RAHM (editors), Lecture Notes in Computer Science, Springer, July 2014, vol. 8574, pp. 16 - 31 [DOI : 10.1007/978-3-319-08590-6_2], <https://hal.inria.fr/hal-01095591>
- [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>
- [91] J.-S. SERENI, J. VOLEC. *A note on acyclic vertex-colorings*, Loria & Inria Grand Est, 2013, <https://hal.archives-ouvertes.fr/hal-00921122>
- [92] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN, A. BOC, V. MAKARENKOV. *A fast compound algorithm for mining generators, closed itemsets, and computing links between equivalence classes*, in "Annals of Mathematics and Artificial Intelligence", 2014, vol. 70, pp. 81 - 105 [DOI : 10.1007/s10472-013-9372-8], <https://hal.inria.fr/hal-01101140>