IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2013

Project-Team ORPAILLEUR

Knowledge discovery, knowledge representation, reasoning

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

RESEARCH CENTER
Nancy - Grand Est

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team ORPAILLEUR

Keywords: Knowledge Discovery, Data Mining, Ontologies, Knowledge Representation, Reasoning

Creation of the Project-Team: 2008 January 01.

1. Members

Research Scientists

Amedeo Napoli [Team leader, CNRS, Senior Researcher, HdR]
Marie-Dominique Devignes [CNRS, Researcher, HdR]
Bernard Maigret [CNRS, Senior Researcher, HdR]
Chedy Raïssi [Inria, Researcher]
David Ritchie [Inria, Senior Researcher, HdR]
Jean-Sébastien Sereni [CNRS, Researcher]
Yannick Toussaint [Inria, Researcher, HdR]

Faculty Members

Adrien Coulet [Univ. Lorraine, Associate Professor]
Nicolas Jay [Univ. Lorraine, Associate Professor]
Jean Lieber [Univ. Lorraine, Associate Professor, HdR]
Jean-François Mari [Univ. Lorraine, Professor, HdR]
Emmanuel Nauer [Univ. Lorraine, Associate Professor]
Malika Smaïl-Tabbone [Univ. Lorraine, Associate Professor]
Mario Valencia [Univ. Paris XIII, Associate Professor, Inria delegation since Sep 2013]

External Collaborators

Florence Le Ber [ENGEES Strasbourg, Professor]
Ioanna Lykourentzou [Centre Henri Tudor, Luxembourg, Post-doctoral Fellow]

Engineers

Jérémie Bourseau [Inria ADT]
Emmanuel Bresso [CNRS, until Aug 2013]
Anisah Ghoorah [Inria, ANR PEPSI project]
Laura Infante Blanco [Inria ADT until Nov 2013]
Luis-Felipe Melo [ANR HYBRIDE and ISTEX Project]
Matthieu Osmuk [Inria, OSEO Innovation Contract, from Jul 2013]
Nicolas Pépin-Hermann [Inria, OSEO Innovation Contract, from Feb 2013]

PhD Students

Mehwish Alam [Inria, OSEO Innovation Grant]
Aleksey Buzmakov [Inria, OSEO Innovation Grant]
Victor Codocedo [Inria, OSEO Innovation Grant]
Sébastien Da Silva [INRA-Inria Grant]
Valmi Dufour-Lussier [Univ. Lorraine, MESR Grant]
Elias Egho [Univ. Lorraine, INCA Grant]
Emmanuelle Gaillard [Univ. Lorraine, MESR Grant]
Ghania Khensous [Univ d'Oran, Algérie, until Dec 2013]
Thomas Meilender [Univ. Lorraine, ATER until Aug 2013]
Gabin Personeni [Univ. Lorraine, MESR Grant]
Mohsen Sayed [Inria, ANR HYBRIDE Grant, from Apr 2013]
My Thao Tang [Univ. Lorraine, ANR Kolflow + Lorraine Region Grant]

Post-Doctoral Fellows

Yasmine Assess [Univ. Lorraine, ATER until Aug 2013]
Aurélie Bertaux [Univ. Lorraine, ATER until Aug 2013]
Melisachew Chekol [Inria, from Jan 2013]
Alice Hermann [Univ. Lorraine, from Jan 2013]
Van-Thai Hoang [Univ. Lorraine, ANR PEPSI Grant, until Dec 2013]

Administrative Assistants

Emmanuelle Deschamps [Inria]
Sylvie Musilli [Univ. Lorraine]

Others

Lucia Martin Reixach [Inria, Internship, from Jun 2013 until Nov 2013]
Camille Sauder [CNRS, from Feb 2013 until Dec 2013]
Renzo Francesco Stanley Cotrozo [Inria, Internship, from May 2013 until Aug 2013]
Niruba Thiagarajan [Inria, Internship, from Jul 2013 until Sep 2013]
Mickaël Zehren [Inria, stagiaire, from Apr 2013 until Aug 2013]

2. Overall Objectives

2.1. Introduction

Knowledge discovery in databases –hereafter KDD– consists in processing a large volume 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. 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 and having a certain knowledge of the task and of the location, the analyst may use his own knowledge but also knowledge on the domain of data for improving the KDD process. Actually, knowledge can be used at each step for improving the KDD process.

Accordingly, a way for the KDD process to take advantage of domain knowledge is to be in connection with ontologies relative to the domain of data, for implementing *knowledge discovery guided by domain knowledge* or KDDK. In the KDDK process, the extracted knowledge units have still "a life" after the interpretation step: they are represented using a knowledge representation formalism to be integrated within an ontology and reused for problem-solving needs. In this way, knowledge discovery is used for extending and updating existing ontologies, showing that knowledge discovery and knowledge representation are complementary tasks, reifying the notion of KDDK.

2.2. Highlights of the Year

For the highlights of the year, we would like to mention the work of Anisah Ghoorah on the KBDOCK system which was accepted in the Database issue of Nucleic Acids Research [6] and as well her paper on "Protein Docking Using Case-Based Reasoning" for the special issue CAPRI [21].

3. Research Program

3.1. From KDD to KDDK

Keywords: knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining

Knowledge discovery in databases is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems. From an operational point of view, the KDD process is performed within a KDD system including databases, data mining modules, and interfaces for interactions, e.g. editing and visualization. The KDD process is based on three main operations: selection and preparation of the data, data mining, and finally interpretation of the extracted units. The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* that are either symbolic or numerical:

- Symbolic methods are based on frequent itemsets search, association rule extraction [108], Formal Concept Analysis and extensions [93].
- Numerical methods are based on higher order stochastic models, namely second-order Hidden Markov Models (HMM2) and Hidden Markov fields (HMRF), which are especially designed for an efficient modeling of space and time [9].

The principle summarizing KDDK can be understood as a process going from complex data units to knowledge units being guided by domain knowledge [104]. Two original aspects can be underlined: (i) the knowledge discovery process is guided by domain knowledge at each step of the process, and (ii) the extracted units are embedded within a knowledge-based system for problem solving purposes.

The KDDK process in the research work of Orpailleur is mainly based on *classification*, which is a polymorphic process involved in modeling, mining, representing, and reasoning tasks. Finally, the KDDK process is intended to feed knowledge-based systems working in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine, and also in the context of semantic web for text mining, information retrieval, and ontology engineering [96], [81].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

Keywords: knowledge discovery, data mining, formal concept analysis, classification, frequent itemset search, association rule extraction, second-order Hidden Markov Models

Classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), 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 [93]. Concept lattices are also called Galois lattices [82].

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 may 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 to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications [119], [118].

Among useful patterns extracted from a database, frequent itemsets are usually thought to unfold "regularities" in the data, i.e. they are the witnesses of recurrent phenomena and they are consistent with the expectations of the domain experts. In some situations however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain. In this way, rare itemsets are related to "exceptions" and thus may convey information of high interest for experts in domains such as biology or medicine [120], [121].

From the numerical point of view, a Hidden Markov Model (HMM2) is a stochastic process aimed at extracting and modeling a sequence of stationary distributions of events. Such models can be used for data mining purposes, especially for spatial and temporal data as they show good capabilities to locate patterns both in time and space domains.

Moreover, stochastic models have been designed to mine temporal sequences having a spatial dimension, for example the succession of land uses in a territory. One main Markovian assumption states that the temporal event succession in a given place depends only on the temporal event successions in neighboring points. By means of stochastic models such as hierarchical hidden Markov models and Markov random fields, it is possible to perform an unsupervised clustering of a spatial territory for discovering "patches" characterized by time and space regularities in their temporal successions. A special effort is currently aimed at designing interactive visualization tools to provide the expert a user-friendly interface.

3.3. Elements on Text Mining

Keywords: knowledge discovery form large collection of texts, text mining, information extraction, document annotation, ontologies

The objective of a text mining process is to extract useful knowledge units from large collections of texts [80], [88]. 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 knowledge units from texts with the help of background knowledge encoded within an ontology (also useful for annotation and relating notions present in texts). Text mining is especially useful in the context of semantic web for ontology engineering [86], [85], [84]. 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. 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".

3.4. Elements on Knowledge Systems and Semantic Web

Keywords: knowledge representation, ontology, description logics, classification-based reasoning, case-based reasoning, semantic web, information retrieval

Usually, people try to take advantage of the web by searching for information (navigation, exploration), and by querying documents using search engines (information retrieval). Then people try to analyze the obtained results, a task that may be difficult and tedious. Semantic web is an attempt for guiding search for information with the help of software agents, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, 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 ontologies are of main importance with respect to the task of setting up semantic web. Thus, there is a need for representation languages for annotating documents, describing the content of documents and giving a semantics to this content. Knowledge representation languages are good candidates for achieving the task: they have a syntax with an associated semantics, and they can be used for retrieving information, answering queries, and reasoning.

Semantic web constitutes a good platform for experimenting ideas on knowledge representation, reasoning, and KDDK. In particular, the knowledge representation language used for designing ontologies is the OWL language, which is based on description logics (DLs [79]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation (i.e. a partial ordering).

The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Furthermore, classification-based reasoning can be extended into case-based reasoning (CBR), which relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem, and possibly memorized for further reuse.

Still in the context of semantic web, research work is also carried on semantic wikis which are web sites for collaborative editing, in which documents can be annotated thanks to semantic annotations and typed relations between wiki pages. Such links provide kind of primitive knowledge units that can be used for guiding information retrieval or knowledge discovery.

Keywords: graph theory, graph mining

4. Application Domains

4.1. Life Sciences

Participants: Yasmine Assess, Emmanuel Bresso, Adrien Coulet, Marie-Dominique Devignes, Elias Egho, Anisah Ghoorah, Nicolas Jay, Bernard Maigret, Amedeo Napoli, Nicolas Pépin-Hermann, Gabin Personeni, David Ritchie, Mohsen Sayed, Malika Smaïl-Tabbone, Yannick Toussaint.

Keywords: knowledge discovery in life sciences, bioinformatics, biology, chemistry, genomics

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, when they exist, solutions bring new research ideas for biologists and for computer scientists as well. Accordingly, the Orpailleur team includes biologists, chemists, and a physician, making Orpailleur a very original EPI at Inria.

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. 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.

4.2. Knowledge Management in Medicine

Participants: Nicolas Jay, Jean Lieber, Thomas Meilender, Amedeo Napoli.

Keywords: knowledge representation, description logics, classification-based reasoning, case-based reasoning, formal concept analysis, semantic web

The Kasimir research project holds on decision support and knowledge management for the treatment of cancer [103]. This is a multidisciplinary research project in which participate researchers in computer science (Orpailleur), experts in oncology ("Institut de Cancérologie de Lorraine Alexis Vautrin" in Vandœuvre-lès-Nancy), Oncolor (a healthcare network in Lorraine involved in oncology), and A2Zi (a company working in Web technologies and involved in several projects in the medical informatics domain, http://www.a2zi.fr/). For a given cancer localization, a treatment is based on a protocol similar to a medical guideline, and is built according to evidence-based medicine principles. For most of the cases (about 70%), a straightforward application of the protocol is sufficient and provides a solution, i.e. a treatment, that can be directly reused. A case out of the 30% remaining cases is "out of the protocol", meaning that either the protocol does not provide a treatment for this case, or the proposed solution raises difficulties, e.g. contraindication, treatment impossibility, etc. For a case "out of the protocol", oncologists try to *adapt* the protocol. Actually, considering the complex case of breast cancer, oncologists discuss such a case during the so-called "breast cancer therapeutic decision meetings", including experts of all specialties in breast oncology, e.g. chemotherapy, radiotherapy, and surgery.

The semantic Web technologies are used and adapted in the Kasimir project since several years [12]. A semantic wiki allowing the management of decision protocols was deployed as an operational system (http://www.oncologik.fr). More precisely, the migration from the static HTML site of Oncolor to a semantic wiki (with limited editing rights and unlimited reading rights) was performed. As a consequence, the editorial chain of the published protocols is more collaborative. A decision tree editor was developed and integrated into this semantic wiki with an export facility to formalized protocols in OWL DL.

4.3. Cooking

Participants: Valmi Dufour-Lussier, Emmanuelle Gaillard, Laura Infante Blanco, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer.

Keywords: cooking, knowledge representation, knowledge discovery, case-based reasoning, semantic wiki

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. Taaable has been at the origin of the ANR CONTINT project Kolflow, whose application domain is WikiTaaable, the semantic wiki of Taaable.

4.4. Agronomy

Participants: Sébastien Da Silva, Florence Le Ber [contact person], Jean-François Mari.

Keywords: simulation, Markov model, Formal Concept Analysis, graph

Sébastien da Silva is working for his PhD thesis in the framework of an Inria-INRA collaboration, which takes place in the INRA research network PAYOTE about landscape modeling. The thesis, supervised both by Claire Lavigne (DR in ecology, INRA Avignon) and Florence Le Ber, is concerned with the characterization and the simulation of hedgerows structures in agricultural landscapes, based on Hilbert-Peano curves and Markov models.

An on-going research work about the representation of peasant knowledge is involved within a collaboration with IRD in Madagascar. Sketches drawn by peasants were transformed into graphs and compared thanks to Formal Concept Analysis [32].

5. Software and Platforms

5.1. Generic Symbolic KDD Systems

5.1.1. The Coron Platform

Participants: Jérémie Bourseau [contact person], Aleksey Buzmakov, Victor Codocedo, Adrien Coulet, Amedeo Napoli, Yannick Toussaint.

Keywords: data mining, frequent itemset, closed itemset, generator, association rule, rare itemset

The Coron platform [117], [101] is a KDD toolkit organized around three main components: (1) Coron-base, (2) AssRuleX, and (3) pre- and post-processing modules. The software was registered at the "Agence pour la Protection des Programmes" (APP) and is freely available (see http://coron.loria.fr). The Coron-base component includes a complete collection of data mining algorithms for extracting itemsets such as frequent itemsets, closed itemsets, generators and rare itemsets. In this collection we can find APriori, Close, Pascal, Eclat, Charm, and, as well, original algorithms such as ZART, Snow, Touch, and Talky-G. AssRuleX generates different sets of association rules (from itemsets), such as minimal non-redundant association rules, generic basis, and informative basis. In addition, the Coron system supports the whole life-cycle of a data mining task and proposes modules for cleaning the input dataset, and for reducing its size if necessary. The Coron toolkit is developed in Java, is operational, and was already used in several research projects.

5.1.2. Orion: Skycube Computation Software

Participant: Chedy Raïssi [contact person].

Keywords: skyline, skycube

This program implements the algorithms described in a research paper published at VLDB 2010 [111]. The software provides a list of four algorithms discussed in the paper in order to compute skycubes. This is the most efficient –in term of space usage and runtime– implementation for skycube computation (see https://github.com/leander256/Orion).

5.2. Stochastic systems for knowledge discovery and simulation

5.2.1. The CarottAge system

Participants: Florence Le Ber, Jean-François Mari [contact person].

Keywords: Hidden Markov Models, stochastic process

The system CarottAge is based on Hidden Markov Models of second order and provides a non supervised temporal clustering algorithm for data mining and a synthetic representation of temporal and spatial data. CarottAge is currently used by INRA researchers interested in mining the changes in territories related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination. CarottAge is also used for mining hydromorphological data. Actually a comparison was performed with three other algorithms classically used for the delineation of river continuum and CarottAge proved to give very interesting results for that purpose [102].

CarottAge is freely available under GPL license (see http://www.loria.fr/~jfmari/App/).

5.2.2. The ARPEnTAge system

Participants: Florence Le Ber, Jean-François Mari [contact person].

Keywords: Hidden Markov Models, stochastic process

ARPEnTAge [1] (for *Analyse de Régularités dans les Paysages: Environnement, Territoires, Agronomie* is a software based on stochastic models (HMM2 and Markov Field) for analyzing spatio-temporal data-bases [107]. ARPEnTAge is built on top of the CarottAge system to fully take into account the spatial dimension of input sequences. It takes as input an array of discrete data in which the columns contain the annual land-uses and the rows are regularly spaced locations of the studied landscape. It performs a Time-Space clustering of a landscape based on its time dynamic Land Uses (LUS). Displaying tools and the generation of Time-dominant shape files have also been defined.

ARPEnTAge is freely available (GPL license) and is currently used by INRA researchers interested in mining the changes in territories related to the loss of biodiversity (projects ANR BiodivAgrim and ACI Ecoger) and/or water contamination. In these practical applications, CarottAge and ARPEnTAge aim at building a partition –called the hidden partition– in which the inherent noise of the data is withdrawn as much as possible. The estimation of the model parameters is performed by training algorithms based on the Expectation Maximization and Mean Field theories. The ARPEnTAge system takes into account: (i) the various shapes of the territories that are not represented by square matrices of pixels, (ii) the use of pixels of different size with composite attributes representing the agricultural pieces and their attributes, (iii) the irregular neighborhood relation between those pixels, (iv) the use of shape files to facilitate the interaction with GIS (geographical information system).

ARPEnTAge and CarottAge were used for mining decision rules in a territory showing environmental issues. They provide a way of visualizing the impact of farmers decision rules in the landscape and revealing new extra hidden decision rules [116].

5.3. KDD in Systems Biology

5.3.1. IntelliGO online

The IntelliGO measure computes semantic similarity between terms from a structured vocabulary (Gene Ontology: GO) and uses these values for computing functional similarity between genes annotated by sets of GO terms [83]. The IntelliGO measure is available on line (http://plateforme-mbi.loria.fr/intelligo/) to be used evaluation purposes. It is possible to compute the functional similarity between two genes, the intra-set similarity value in a given set of genes, and the inter-set similarity value for two given sets of genes.

5.3.2. WAFOBI : KNIME nodes for relational mining of biological data

KNIME (for "Konstanz Information Miner") is an open-source visual programming environment for data integration, processing, and analysis. KNIME includes a rich library of data manipulation tools (import, export) and several mining algorithms which operate on a single data matrix (decision trees, clustering, frequent itemsets, association rules...). The KNIME platform aims at facilitating the data mining experiment settings as many tests are required for tuning the mining algorithms. The evaluation of the mining results is also an important issue and its configuration is made easier.

Various KNIME nodes were developed for supporting relational data mining using the ALEPH program (http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.pl). These nodes include a data preparation node for defining a set of first-order predicates from a set of relation schemes and then a set of facts from the corresponding data tables (learning set). A specific node allows to configure and run the ALEPH program to build a set of rules. Subsequent nodes allow to test the first-order rules on a test set and to perform configurable cross validations.

5.3.3. MOdel-driven Data Integration for Mining (MODIM)

Participants: Marie-Dominique Devignes [contact person], Malika Smaïl-Tabbone.

[1] http://www.loria.fr/~jfmari/App/

The MODIM software (MOdel-driven Data Integration for Mining) is a user-friendly data integration tool which can be summarized along three functions: (i) building a data model taking into account mining requirements and existing resources; (ii) specifying a workflow for collecting data, leading to the specification of wrappers for populating a target database; (iii) defining views on the data model for identified mining scenarios. A version of the software was declared through Inria APP procedure in December, 2010.

Although MODIM is domain independent, it was used so far for biological data integration in various internal research studies. MODIM was also used for organizing data about non ribosomal peptide syntheses. The sources can be downloaded at https://gforge.inria.fr/projects/modim/.

5.4. Knowledge-Based Systems and Semantic Web Systems

5.4.1. The Kasimir System for Decision Knowledge Management

Participants: Nicolas Jay, Jean Lieber [contact person], Amedeo Napoli, Thomas Meilender.

Keywords: classification-based reasoning, case-based reasoning, decision knowledge management, knowledge edition, knowledge base maintenance, semantic portal

The objective of the Kasimir system is decision support and knowledge management for the treatment of cancer. A number of modules have been developed within the Kasimir system for editing treatment protocols, visualization, and maintenance. Kasimir is developed within a semantic portal, based on OWL. KatexOWL (Kasimir Toolkit for Exploiting OWL Ontologies, http://katexowl.loria.fr) is developed in a generic way and is applied to Kasimir. In particular, the user interface EdHibou of KatexOWL is used for querying the protocols represented within the Kasimir system (see [17] where an an extension of Kasimir for multi-viewpoint case-based reasoning is presented).

CabamakA (case base mining for adaptation knowledge acquisition) is a module of the Kasimir system. This system performs case base mining for adaptation knowledge acquisition and provides information units to be used for building adaptation rules. Actually, the mining process in CabamakA is based on a frequent close itemset extraction module from the Coron platform (see §5.1.1).

The Oncologik system [12] is a collaborative editing tool aiming at facilitating the management of medical guidelines (http://www.oncologik.fr/). Based on a semantic wiki, it allows the acquisition of formalized decision knowledge. Oncologik also includes a graphical decision tree editor called KcatoS.

5.4.2. Taaable: a system for retrieving and creating new cooking recipes by adaptation

Participants: Valmi Dufour-Lussier, Emmanuelle Gaillard, Laura Infante Blanco, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer [contact person].

Keywords: knowledge acquisition, ontology engineering, semantic annotation, case-based reasoning, hierarchical classification, text mining

Taaable [69] is a system whose objectives are to retrieve textual cooking recipes and to adapt these retrieved recipes whenever needed. Suppose that someone is looking for a “leek pie” but has only an “onion pie” recipe: how can the onion pie recipe be adapted?

The Taaable system combines principles, methods, and technologies such as case-based reasoning (CBR), ontology engineering, text mining, text annotation, knowledge representation, and hierarchical classification. Ontologies for representing knowledge about the cooking domain, and a terminological base for binding texts and ontology concepts, were built from textual web resources. These resources are used by an annotation process for building a formal representation of textual recipes. A CBR engine considers each recipe as a case, and uses domain knowledge for reasoning, especially for adapting an existing recipe w.r.t. constraints provided by the user, holding on ingredients and dish types.

The Taaable system is available on line since 2008 at http://taaable.fr. A new version of Taaable was implemented using Tuuurbine, a generic ontology-guided CBR engine based on semantic web technologies (see Section 5.4.3). BeGoood (see Section 5.4.4), a generic system for managing non-regression tests on knowledge bases, is also plugged for acquiring test sets. When the Taaable system returns answers to a query, the user may evaluate the relevance of the answers. Currently, user feedback is collected using BeGoood and will be used in the future to run tests when the knowledge exploited by the CBR system evolves. The objective is to ensure that the knowledge base evolution does not affect the quality of answers given by the CBR system.

5.4.3. *Tuuurbine: a generic ontology guided case-based inference engine*

Participants: Laura Infante Blanco, Jean Lieber, Emmanuel Nauer [contact person].

Keywords: case-based reasoning, inference engine, knowledge representation, ontology engineering, semantic web

The experience acquired since 5 years with the Taaable system conducted to the creation of a generic cased-based reasoning system, whose reasoning procedure is based on a domain ontology. This new system, called Tuuurbine (http://tuuurbine.loria.fr/), takes into account the retrieval step, the case base organization, but also an adaptation procedure which is not addressed by other generic case-based reasoning tools. Moreover, Tuuurbine is built over semantic web standards allowing to be connected to the web of data. The domain knowledge is represented in an RDF store, which can be interfaced with a semantic wiki, for collaborative edition and management of the knowledge involved in the reasoning system (cases, ontology, adaptation rules). The development of Tuuurbine was supported by an Inria ADT funding until October 2013.

5.4.4. *BeGoood: a generic system for managing non-regression tests on knowledge-bases*

Participants: Laura Infante Blanco, Emmanuel Nauer [contact person].

Keywords: tests, non-regression, knowledge evolution

BeGoood [67] is a system allowing to define test plans, independent of any application domain, and usable for testing any system answering queries by providing results in the form of sets of strings. BeGoood provides all the features usually found in test systems, such as tests, associated queries, assertions, and expected result sets, test plans (sets of tests) and test reports. The system is able to evaluate the impact of a system modification by running again test plans and by evaluating the assertions which define whether a test fails or succeeds. The main components of BeGoood are (1) the "test database" that stores every test artifacts, (2) the "remote query evaluator" which evaluates test queries, (3) the "assertion engine" which evaluates assertions over the expected and effective query result sets. and finally (4) the "REST API" which offers the test functionalities as web services.

BeGoood is available under a AGPL license on github [2]. BeGoood is used to manage the non-regression of the Taaable system (see Section 5.4.2) when the knowledge base used by the CBR system is modified.

5.4.5. *Revisor: a library of revision operators and revision-based adaptation operators*

Participants: Valmi Dufour-Lussier, Alice Hermann, Florence Le Ber, Jean Lieber [contact person], Emmanuel Nauer, Gabin Personeni.

Keywords: belief revision, adaptation, revision-based adaptation, case-based reasoning, inference engines, knowledge representation

Revisor is a library of inference engines dedicated to belief revision and to revision-based adaptation for case-based reasoning [60]. It is open source, under a GPL license and available on the web (http://revisor.loria.fr/). It gathers several engines developed during the previous years, for various knowledge representation formalisms (propositional logic—with or without the use of adaptation knowledge [65]—conjunction of linear constraints, and qualitative algebras [3]). Some of these engines are already used in the Taaable system. Current developments on Revisor aim at defining new engines in other formalisms.

[2] https://github.com/kolflow/begoood

6. New Results

6.1. The Mining of Complex Data

Participants: Mehwish Alam, Aleksey Buzmakov, Melisachew Chekol, Victor Codocedo, Adrien Coulet, Elias Egho, Nicolas Jay, Florence Le Ber, Ioanna Lykourentzou, Luis-Felipe Melo, Amedeo Napoli, Chedy Raïssi, Mohsen Sayed, My Thao Tang, Mohsen Sayed, Yannick Toussaint.

Keywords: formal concept analysis, relational concept analysis, pattern structures, frequent itemset, association rule, graph mining, sequence mining, skyline

Formal Concept Analysis, together with itemset search and association rule extraction, are suitable symbolic methods for KDDK, that may be used for real-sized applications. Global improvements are carried 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 biological or chemical data or textual documents, involving objects with multi-valued attributes (e.g. domains or intervals), n-ary relations, sequences, trees and graphs.

6.1.1. FCA and variations: RCA and Pattern Structures

There are a few extensions of FCA for handling contexts involving complex data formats, e.g. graphs or relational data. Among them, Relational Concept Analysis (RCA) is a process for analyzing objects described both by binary and relational attributes [10]. The RCA process takes as input a collection of contexts and of inter-context relations, and yields a set of lattices, one per context, whose concepts are linked by relations. RCA has an important role in KDDK, especially in text mining [86], [85].

Another extension of FCA is based on Pattern Structures (PS) [92], which allows to build a concept lattice from complex data, e.g. nominal, numerical, and interval data. In [100], pattern structures are used for building a concept lattice from interval data. Since then, we worked on a some experiments involving pattern structures, namely sequence mining [41], information retrieval [48] and functional dependencies [38]. one of the next step is the adaptation of pattern structures to graph mining. Moreover, the notion of similarity between objects is also closely related to pattern structures [99]: two objects are similar as soon as they share the same attributes (binary case) or attributes with similar values or the same description (at least in part). Combination of similarity and pattern structures is also under study, in particular for solving information retrieval and annotation problems.

Finally, there is also an on-going work relating FCA and semantic web. This work focuses on the classification within a concept lattice of the answers returned by SPARQL queries [37], [47], [46], [44]. The concept lattice is then used as an index for navigating and ranking the answers w.r.t. their content and interest for a given objective.

6.1.2. Advances in mining complex data: sequences and healthcare trajectories

Sequence data is widely used in many applications. Consequently, mining sequential patterns and other types of knowledge from sequence data has become an important data mining task. The main emphasis has been on developing efficient mining algorithms and effective pattern representation. The most frequent sequences generally provide a trivial information. When analyzing the set of frequent sequences with a low minimum support, the user is overwhelmed by millions of patterns. In our recent work, the general idea is to extract patterns whose characteristic on a given measure such as the support strongly deviates from its expected value under a null model. The frequency of a pattern is considered as a random variable, whose distribution under the null model has to be calculated or approximated. Then, the significance of the pattern is assessed through a statistical test that compares the expected frequency under the null model to the observed frequency. One of the key-points of this family of approaches is to choose an appropriate null model. It will ideally be a trade-off between adjustment to the data and simplicity: the model should capture some characteristics of the data, to integrate prior knowledge, without overfitting, to allow for relevant patterns discovery. We introduced a

rigorous and efficient approach to mine statistically significant, unexpected patterns in sequences of itemsets. Experiments on sequences of replays of a video game demonstrated the scalability and the efficiency of the method to discover unexpected game strategies. This work was successfully published as an international conference paper [8].

Other work on sequences is in concern with patient trajectories, i.e. the "path" of a patient during its illness. With the increasing burden of chronic illnesses, administrative health care databases hold valuable information that could be used to monitor and assess the processes shaping the trajectory of care of chronic patients. In this context, temporal data mining methods are promising tools, though lacking flexibility in addressing the complex nature of medical events. In a set of recent works with Elias Egho, a PhD candidate, we present new algorithms to extract patient trajectory patterns with different levels of granularity by relying on external taxonomies [52]. Our algorithms rely on the general FCA framework to formalize the general notion of multidimensional healthcare trajectories. We also continued working on the complex notion of sequences or trajectory similarity measures. We show the interest of our approaches with the analysis of trajectories of care for colorectal cancer using data from the French healthcare information system (see also [41]).

6.1.3. KDDK in Text Mining

Ontologies help software and human agents to communicate by providing shared and common domain knowledge, and by supporting various tasks, e.g. problem-solving and information retrieval. In practice, building an ontology depends on a number of "ontological resources" having different types: thesaurus, dictionaries, texts, databases, and ontologies themselves. We are currently working on the design of a methodology and the implementation of a system for ontology engineering from heterogeneous ontological resources [58]. This methodology is based on both FCA and RCA, and was previously successfully applied in contexts such as astronomy and biology. In the framework of the ANR Hybride project (see 8.2.1.2), an engineer is implementing a robust system based on these previous research results, for preparing the way to new research directions involving trees and graphs.

6.2. KDDK in Life Sciences

Participants: Yasmine Assess, Emmanuel Bresso, Adrien Coulet, Marie-Dominique Devignes, Anisah Ghoorah, Bernard Maigret, Amedeo Napoli, Gabin Personeni, David Ritchie, Mohsen Sayed, Malika Smaïl-Tabbone, My Thao Tang, Mohsen Sayed, Yannick Toussaint.

The 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 health care, 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 the Life Sciences is developed in one main direction which is in concern with the use of bio-ontologies to improve KDDK but also information retrieval, access to the so-called "Linked Open Data" and data integration.

6.2.1. Using ILP for the characterization and prediction of drug side-effect profiles

Inductive Logic Programming (ILP) is a learning method which allows expressive representation of the data and produces explicit first-order logic rules [89]. We applied ILP for understanding drug side-effets. Indeed, late appearance of adverse side effets during clinical trials constitute the main reason for stopping the drug development process which is very costly [1]. Improving our ability to understand drug side effects is necessary to reduce this inconvenience. Moreover, it can contribute to design safer drugs and anticipate the appearance of yet unreported side effects of approved drugs. Today, most investigations deal with prediction of single side effects and overlook possible combinations.

In our study, drug annotations are collected from the SIDER and DrugBank databases. Terms describing individual side effects reported in SIDER are clustered with the IntelliGO semantic similarity measure into term clusters (TCs) [83]. Maximal frequent itemsets are extracted from the resulting $drug \times TC$ binary table, leading to the identification of what we call side-effect profiles (SEPs). A SEP is defined as the longest combination of TCs which are shared by a significant number of drugs. Frequent SEPs are explored on the basis of integrated drug and target descriptors using two machine learning methods: decision-trees and ILP. Learning efficiency is evaluated by cross-validation and direct testing with new molecules. Comparison of the two methods shows that the ILP displays a greater sensitivity than decision trees. Although both methods yield explicit models, ILP is able to exploit not only drug properties but also background knowledge, thereby producing rich and expressive rules.

6.2.2. *Functional classification of genes*

The IntelliGO measure computes semantic similarity between genes in taking into account domain knowledge in Gene Ontology (GO) [83]. IntelliGO is used for functional clustering of a set of genes, i.e. based on functional annotations of these genes. For example, a gene set of interest may include genes showing the same expression profile.

A functional clustering method based on IntelliGO was tested on four benchmarking datasets consisting of biological pathways (KEGG database) and functional domains (Pfam database) [90]. A follow-up of this study was motivated by the fact that the IntelliGO measure, like most of the biological similarity measures, does not verify "triangle inequality" and thus is not a mathematical distance. Interestingly, specific spectral clustering techniques can be used for improving the clustering of the objects for which exists a pairwise (dis-)similarity matrix [115], [125]. Spectral clustering techniques make use of the eigenvalues of this (dis-)similarity matrix to perform dimension reduction before clustering in fewer dimensions. We have conducted a comparative and large-scale gene clustering evaluation using the IntelliGO measure and reference sets. Our results showed an improvement of the clustering quality with "constant-shift spectral clustering" [63].

6.2.3. *Analysis of biomedical data annotated with ontologies*

Annotating data with concepts of an ontology is a common practice in the biomedical domain. Resulting annotations define links between data and ontologies that are key for data exchange, data integration and data analysis. Since 2011, we collaborate with the National Center for Biomedical Ontologies (NCBO) to develop a large repository of annotations named the NCBO Resource Index [98]. This repository contains annotations of 36 biomedical databases annotated with concepts of more than 200 ontologies of the BioPortal [3]. In 2012, we compared the annotations of a database of biomedical publications (Medline) with two databases of scientific funding (Crisp and ResearchCrossroads) to profile disease research [105]. One main challenge remains to develop a knowledge discovery approach able to mine correlations between annotations based on BioPortal ontologies, i.e. is it possible to discover interesting knowledge units within these annotations?

In 2013, we proposed an adaptation of FCA techniques, namely pattern structures, to explore the annotations of biomedical databases [2]. We considered documents of biomedical databases annotated with sets of ontological concepts as objects in a pattern structure. Corresponding annotations have been classified according to several dimensions, where a dimension is related to a particular aspect of domain knowledge. Then, the pattern structure formalism was applied to classify these annotations, allowing to discover correlations between annotations but also lacks of completion in the annotations that could be fixed afterward. This adaptation of pattern structures opens many perspectives in term of ontology reengineering and knowledge discovery.

In another context, a related work was carried out in the Kolflow project (see 8.2.1.4). We proposed an interactive environment based on Formal Concept Analysis which makes possible a simultaneous enrichment of semantic annotations of medical texts and of the ontology of medical domain [66], [59].

6.2.4. *Analysis and interpretation of sequential patterns with Linked Open Data*

[3] http://bioportal.bioontology.org/

Linked Data is a set of principles and technologies that rely on the architecture of the Web (URIs and links) to share, model and integrate data. The basic idea is that data objects (e.g., a surgical procedure) are identified by web addresses (URIs), and the information attached to these objects are represented through links to values or other URIs representing other objects.

Considering the potential development and availability of biomedical Linked Data, we investigated it as a source of additional information to support the interpretation of the results of a data mining process, such as sequential pattern discovery. We developed a system using several linked data endpoints to collect descriptive dimensions about the items that constitute sequential patterns. These dimensions are used to automatically classify with Formal Concept Analysis the extracted patterns, thus generating a structure that can support exploration and navigation into the results of the data mining step [55].

6.3. Structural Systems Biology

Participants: Marie-Dominique Devignes, Anisah Ghoorah, Van-Thai Hoang, Bernard Maigret, David Ritchie, Malika Smaïl-Tabbone.

Keywords: bioinformatics, chemistry, docking, knowledge discovery, screening, systems biology

Structural systems biology aims to describe and analyze the many components and interactions within living cells in terms of their three-dimensional (3D) molecular structures. We are currently developing advanced computing techniques for molecular shape representation, protein-protein docking, protein-ligand docking, high-throughput virtual drug screening, and knowledge discovery in databases dedicated to protein-protein interactions.

6.3.1. Accelerating protein docking calculations using graphics processors

We have recently adapted the *Hex* protein docking software [113] to use modern graphics processors (GPUs) to carry out the expensive FFT part of a docking calculation [114]. Compared to using a single conventional central processor (CPU), a high-end GPU gives a speed-up of 45 or more. This software is publicly available at http://hex.loria.fr. A public GPU-powered server has also been created (http://hexserver.loria.fr) [106]. The docking server has performed some 14,000 docking runs during 2013.

Our docking work has facilitated further developments on modeling the assembly of multi-component molecular structures using a particle swarm optimization technique [123], and on modeling protein flexibility during docking [122]. In 2013, in collaboration with the Nano-D team at Inria Grenoble, we developed a new docking algorithm called "DockTrina" [31], which can rapidly model trimers of protein structures by combining multiple pair-wise docking results from *Hex*. We also used *Hex* successfully to model a challenging protein complex containing water molecules at the protein-protein interface [29].

6.3.2. KBDOCK: Protein docking using Knowledge-Based approaches

In order to explore the possibilities of using structural knowledge of protein-protein interactions, Anisah Ghoorah recently developed the KBDOCK system as part of her doctoral thesis project [95]. KBDOCK is available at http://kbdock.loria.fr. KBDOCK combines coordinate data from the Protein Data Bank [87] with the Pfam protein domain family classification [91] in order to describe and analyze all known protein-protein interactions for which the 3D structures are available. We have demonstrated the utility of KBDOCK [94] for template-based docking using 73 complexes from the Protein Docking Benchmark [97]. We recently presented results obtained using KBDOCK at the CAPRI conference on protein docking in Utrecht [21]. In 2013, we updated KBDOCK with the latest data from Pfam and the Protein Data Bank. An article describing the new version of KBDOCK was accepted by the Database Issue of Nucleic Acids Research [6].

6.3.3. Kpax: A new algorithm for protein structure alignment

We have developed a new protein structure alignment approach called Kpax [112]. The approach exploits the fact that each amino acid residue has a carbon atom with a highly predictable tetrahedral geometry. This allows the local environment of each residue to be transformed into a canonical orientation, thus allowing easy comparison between the canonical orientations of residues within pairs of proteins using a novel scoring function based on Gaussian overlaps. The overall approach is two or three orders of magnitude faster than most contemporary protein structure alignment algorithms, while still being almost as accurate as the state-of-the-art TM-Align approach [124]. The Kpax program is available at http://kpax.loria.fr/. The Kpax program is now used heavily behind the scenes in the new KBDOCK web server [6] to find structural templates for docking which might be beyond the reach of sequence-based homology modeling approaches.

6.3.4. gEMpicker and gEMfitter: GPU-accelerated tools for cryo-electron microscopy

Solving the structures of large protein assemblies is a difficult and computationally intensive task. Multiple two-dimensional (2D) images must be processed and classified to identify protein particles in different orientations. These images may then be averaged and stacked to deduce the three-dimensional (3D) structure of a protein. In order to help accelerate the first of these tasks we have recently developed a novel and highly parallel algorithm called "gEMpicker" which uses multiple graphics processors to detecting 2D particles in cryo-electron microscopy images [112]. We have also developed a 3D shape matching algorithm called "gEMfitter" which also exploits graphics processors, and which will provide a useful tool for the final 3D assembly step [112]. Both programs have been made publicly available at http://gem.loria.fr/.

6.3.5. DOVSA: Developing new algorithms for virtual screening

In 2010, Violeta Pérez-Nueno joined the Orpailleur team thanks to a Marie Curie Intra-European Fellowship (IEF) award to develop new virtual screening algorithms (DOVSA). The aim of this project was to advance the state of the art in computational virtual drug screening by developing a novel consensus shape clustering approach based on spherical harmonic (SH) shape representations [110]. As a continuation of this project, and in collaboration with colleagues from the University of Bari in Italy, we recently published a review on drug discovery relating to the GPCR receptor proteins [15]. We also published a book chapter describing the ParaFit program for fast spherical harmonic shape matching [70].

6.4. Around the Taaable research project

Participants: Valmi Dufour-Lussier, Emmanuelle Gaillard, Laura Infante Blanco, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer.

Keywords: knowledge representation, description logics, classification-based reasoning, case-based reasoning, belief revision, semantic web

The Taaable project [69] (http://taaable.fr) has been originally created as a challenger of the Computer Cooking Contest (ICCBR Conference). A candidate to this contest is a system whose goal is to solve cooking problems on the basis of a recipe book (common to all candidates), where each recipe is a shallow XML document with an important plain text part. The size of the recipe book (about 1500 recipes) prevents from a manual indexing of recipes: this indexing is performed using semi-automatic techniques.

Beyond its participation to the CCCs, the Taaable project aims at federating various research themes: case-based reasoning (CBR), information retrieval, knowledge acquisition and extraction, knowledge representation, minimal change theory, ontology engineering, semantic wikis, text-mining, etc. CBR is used to perform adaptation of recipe to user constraints. The reasoning process uses a cooking domain ontology (especially hierarchies of classes) and adaptation rules. The knowledge base used by the inference engine is encoded within a semantic wiki, which contains the recipes, the domain ontology, and adaptation rules.

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 for the adaptation of recipe preparations [3], using one of the engines included in the library Revisor (cf. § 5.4.5).

As acquiring knowledge from experts is costly, a new approach was proposed to allow a CBR system to use partially reliable, non expert, knowledge from the Web for reasoning [68] [5]. This approach is based on a meta-knowledge model to manage knowledge reliability. This model represents notions such as belief, trust, reputation and quality, as well as their relationships and rules to evaluate 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, ensuring the quality of results.

6.5. Some results in graph theory

Participants: Amedeo Napoli, Chedy Raïssi, Jean-Sébastien Sereni, Mario Valencia.

Keywords: graph theory, extremal graph theory, coloring, clustering

6.5.1. Structural and extremal graph theory

Regarding graph coloring, a conjecture of Gera, Okamoto, Rasmussen and Zhang on set coloring was solved. A *set coloring* of a graph $G = (V, E)$ is a function $c : V \to \{1, ..., k\}$ such that whenever u and v are adjacent vertices, it holds that $\{c(x) : x \text{ neighbor of } u\} \neq \{c(x) : x \text{ neighbor of } v\}$. In other words, there must be at least one neighbor of u that has a color not assigned to a neighbor of v, or *vice-versa*. The smallest k such that G admits a set coloring is the *set coloring number* $\chi_s(G)$. We confirmed the conjecture by proving that $\chi_s(G) \geq \lceil \log_2 \chi(G) \rceil + 1$, where $\chi(G)$ is the (usual) chromatic number of G. This bound is tight.

Works have been started on a 12-year-old conjecture by Heckman and Thomas about the fractional chromatic number of graphs with no triangles and maximum degree at most 3. This conjecture is actually a natural generalization of a fact established by Staton in 1979. Heckman and Thomas posits that in every graph with no triangles, maximum degree at most 3 and arbitrary weights on the vertices, there exists an independent set of weight at least $5/14$ times the total weight of the graph.

Regarding extremal graph theory, two results have been obtained. The first one deals with permutation snarks, while the second one reads as follows.
For every 3-coloring of the edges of the complete graph on n vertices, there is a color c and a set X of 4-vertices such that at least $2n/3$ vertices are linked to a vertex in X by an edge of color c.
This theorem is motivated by a conjecture of Erdős, Faudree, Gould, Gyárfás, Rousseau and Schelp from 1989, which asserts that X can be of size 3 only. However, they were only able to prove that X can be of size 22. Recently, Rahil Baber and John Talbot managed to build upon our work in a very nice article: adding a new idea to our argument, they managed to confirm the conjecture.

6.5.2. Graph theory and other fields

Interactions of graph theory with other topics (theoretical computer science, number theory, group theory, sociology and chemistry) have been considered. Most of them are still in progress and some are published. For instance, regarding distributed computing, the purpose of our work was to question the global knowledge each node is assumed to start with in many distributed algorithms (both deterministic and randomized). More precisely, numerous sophisticated local algorithm were suggested in the literature for various fundamental problems. Noticeable examples are the MIS algorithms and the $(\Delta + 1)$-coloring algorithms. Unfortunately, most known local algorithms are *non-uniform*, that is, they assume that all nodes know good estimations of one or more global parameters of the network, e.g., the number of nodes n. Our work provides a rather general method for transforming a non-uniform local algorithm into a uniform one. Furthermore, the resulting algorithm enjoys the same asymptotic running time as the original non-uniform algorithm. Our method applies to a wide family of both deterministic and randomized algorithms. Specifically, it applies to almost all of the state of the art non-uniform algorithms regarding MIS and Maximal Matching, as well as to many results concerning the coloring problem.

6.5.3. Other aspects on graph coloring and clustering

Since September 2013, Mario Valencia has obtained a one year invitation (namely Inria "Délégation") for working at Inria Nancy – Grand Est, in the Orpailleur team, on graph theoretical aspects and data clustering. This research work consists in studying the modular decomposition techniques on the threshold graphs issues of the clustering process. More precisely, this study relies on families of graphs having a "good" decomposition as cographs and chordal graphs, and then, and on the analysis of the adaptation of these two families of graphs within a clustering activity.

Other research dimensions are dealing with algorithmic aspects of some variations of the classical graph coloring problem.

- Packing colorings of graphs where we need to color the vertices of a graph in such a way that vertices having a same color c should be at a distance at least equal to $c+1$ in the graph. With P. Torres, a postdoc student, we have obtained some upper bounds for the packing chromatic number of hypercubes graphs of dimension n, denoted by Q_n, and we have computed exactly this parameter for this family of graph for $n = 6, 7, 8$, extending previous results known for $n = 2, 3, 4, 5$ [35].
- (k, i)-coloring of graphs, which is a generalization of a k-tuple coloring of graphs: given positive integers k and i, we want to affect to each vertex a k-set of colors such that the intersection of the k-sets affected to adjacent vertices has cardinality at most equal to i. With F. Bonomo, I. Koch, and G. Duran, we have found a linear time algorithm for this problem on cycles and cacti graphs. Moreover, we have obtained an interesting equivalence between this problem on complete graphs and a problem on weighted binary codes.
- b-coloring of graphs, where we need to color the vertices of a graph in such a way that in each color class j there exists at least one vertex x_j adjacent to at least one vertex in all the other color classes. The goal of this problem is to maximize the number of colors under such a constraint (i.e. the b-chromatic number of a graph). With F. Bonomo, O. Schaudt and M. Stein, we have shown that b-coloring is NP-hard on co-bipartite graphs and polytime solvable on tree-cographs [77].

7. Bilateral Contracts and Grants with Industry

7.1. The BioIntelligence Project

Participants: Mehwish Alam, Yasmine Assess, Aleksey Buzmakov, Melisachew Chekol, Adrien Coulet, Marie-Dominique Devignes, Amedeo Napoli [contact person], Nicolas Pépin-Hermann, Malika Smaïl-Tabbone.

The objective of the "BioIntelligence" project is to design an integrated framework for the discovery and the development of new biological products. This framework takes into account all phases of the development of a product, from molecular to industrial aspects, and is intended to be used in life science industry (pharmacy, medicine, cosmetics, etc.). The framework has to propose various tools and activities such as: (1) a platform for searching and analyzing biological information (heterogeneous data, documents, knowledge sources, etc.), (2) knowledge-based models and process for simulation and biology in silico, (3) the management of all activities related to the discovery of new products in collaboration with the industrial laboratories (collaborative work, industrial process management, quality, certification). The "BioIntelligence" project is led by "Dassault Systèmes" and involves industrial partners such as Sanofi Aventis, Laboratoires Pierre Fabre, Ipsen, Servier, Bayer Crops, and two academics, Inserm and Inria. An annual meeting of the project usually takes place in Sophia-Antipolis at the beginning of July.

Two theses related to "BioIntelligence" are currently in preparation within the Orpailleur team. A first thesis is related to the mining of complex biological data using FCA and RCA techniques [37], [44]. The objective is to take advantage of Linked Open data in biology for helping the biologist querying complex data. There are needs to integrate data and knowledge from several web resources. Practical experiments will be led on biological data (clinical trials data and cohort data) also in accordance with ontologies lying at the NCBO BioPortal.

A second thesis is based on an extension of FCA involving Pattern Structures on complex data such as sequences and graphs [42], [41]. The idea is to extend the formalism of pattern structures to these complex data for being able to classify complex structures such as patient trajectories or molecular structures. The classification results (e.g. concept lattices) are expected to help practitioners in information retrieval tasks and specific problem solving.

7.2. The Quaero Project

Participants: Victor Codocedo [contact person], Ioanna Lykourentzou, Amedeo Napoli.

The Quaero project (http://www.quaero.org) is a program aimed at promoting research and industrial innovation on technologies for automatic analysis and classification of multimedia and multilingual documents. The partners collaborate on research and the realization of advanced demonstrators and prototypes of innovating applications and services for access and usage of multimedia information, such as spoken language, images, video and music.

In this framework, the Orpailleur team is working on information retrieval, document annotation and recommendation. The objective is to define methods and algorithms for achieving these complex tasks, based onKDDK techniques and especially the FCA technology.

A thesis is in preparation in the context of the Quaero project, where information retrieval and document annotation are especially studied, namely information retrieval guided by domain knowledge and classification of documents w.r.t. their annotations using FCA and pattern structures [48]. In addition, a related work was carried out on the reengineering of relational data within a concept lattice [58].

8. Partnerships and Cooperations

8.1. International Initiatives

8.1.1. Participation In International Programs

8.1.1.1. Facepe Inria Project: CM2ID

Participants: Amedeo Napoli [contact person], Chedy Raïssi.

This research project called “Combining Numerical and Symbolical Methods for the Classification of Multi-valued and Interval Data (CM2ID)” involves the Orpailleur Team at Inria NGE, AxIS at Inria Rocquencourt (Yves Lechevallier) and the computer science laboratory of the University of Recife (Prof. Francisco de A.T. de Carvalho). The project aims at developing and comparing classification and clustering algorithms for 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.

Finally, a workshop was organized in April 2013 , namely the “French-Brazilian Workshop on Numerical and Symbolic Methods of Data Analysis -WFB2013” (http://www.cin.ufpe.br/~wfb2013/).

8.1.1.2. Fapemig Inria Project: IKMSDM

Participants: Amedeo Napoli [contact person], Chedy Raïssi.

This Fapemig – Inria research project, called “Incorporating knowledge models into scalable data mining algorithms” involves researchers at Universidade Federal de Minas Gerais in Belo Horizonte –a group led by Prof. Wagner Meira– and the Orpailleur team at Inria Nancy Grand Est. In this project we are interested in the mining of large amount of data and we target two relevant application scenarios where such issue may be observed. The first one is text mining, i.e. extracting knowledge from texts and document categorization. The second application scenario is graph mining, i.e. determining relationship-based patterns and use these relations to perform classification tasks. In both cases, the computational complexity is large either because the high dimensionality of the data or the complexity of the patterns to be mined. Loïc Cerf from UFMG visited the Orpailleur team in January 2013 while Chedy Raïssi visited UFMG in May 2013.

8.1.1.3. Pronex Brasilia

Participant: Bernard Maigret [contact person].

In this research project, the goal is to identify, using virtual screening techniques that we developed, new compounds against tropical diseases (e.g. trypanosome, dengue and mycosis) in collaboration with several Brazilian laboratories among which the Department of Biology at University of Brasilia, together with the Harmonic Pharma start-up. Through this collaboration, several PhD and postdocs came to the lab for one year training with our home-developed virtual screening engine (VSM-G). This project is in part supported by the Brazilian CNPq agency. Fruitful results were already obtained leading to several papers in preparation and patents. These patents concern the discovery of new putative treatment of strong mycosis due to fungi particularly virulent in South America. These patents were funded by the University of Brasilia, Embrapa and Harmonic Pharma.

8.1.1.4. International collaborations in Mining complex data

Participants: Mehwish Alam, Aleksey Buzmakov, Melisachew Chekol, Victor Codocedo, Adrien Coulet, Elias Egho, Ioanna Lykourentzou, Amedeo Napoli [contact person], Chedy Raïssi, Jean-Sébastien Sereni, Mario Valencia.

8.1.1.4.1. PICS CNRS CAdOE

A collaboration involves the Orpailleur team, "Université du Québec à Montréal" (UQAM) in Montréal with Prof. Petko Valtchev and Laboratoire LIRMM in Montpellier with Prof. Marianne Huchard. This collaboration is supported by a CNRS PICS project (2011-2014), which is called "Concept Analysis driving Ontology Engineering" and abbreviated in "CAdOE". The research work within this project is aimed at defining and implementing a semi-automatic methodology supporting ontology engineering based on the joint use of Formal Concept Analysis (FCA) and Relational Concept Analysis (RCA). This year the work was mainly focused on RCA and some important papers were published [33], [57].

8.1.1.4.2. Miscellaneous

- An on-going collaboration involves the Orpailleur team and Sergei Kuznetsov at Higher School of Economics in Moscow (HSE). Amedeo Napoli visited HSE laboratory in March 2013 (with the support of HSE) and met Sergei Kuznetsov several times during the year. In addition, Alexey Neznanov from HSE Moscow visited the Orpailleur team in May 2013 while Dmitry Ignatov visited the visited the Orpailleur team in September 2013.

  These visits were the occasion of preparing a publications. Moreover, Sergei Kuznetsov and Amedeo Napoli, together with Claudio Carpineto organized a workshop related to the ECIR Conference in Moscow in March 2013 on "Formal Concept Analysis meets Information Retrieval" (http://www.hse.ru/en/org/hse/fcair).
- A so-called AGAUR Project funded by UPC Barcelona involves Amedeo Napoli and Jaume Baixeries who is an Associate Professor at UPC Barcelona (Universitat Politècnica de Catalunya). Both researchers have worked, jointly with Mehdi Kaytoue, on the characterization of functional dependencies in many-valued data with FCA and pattern structures [38].
- A PHC Zenon project (Cyprus) with Florent Domenach, associated professor at the University of Nicosia in Cyprus was finished at the end of last year. This project was entitled "Knowledge Discovery for Complex Data in Formal and Relational Concept Analysis" (KD4CD) and is aimed at studying and combining different types of classification process in the framework of FCA. As a result of this collaboration, some papers were published this year, among which one at the ICFCA Conference in Dresden [49], [61].
- A PHC Proteus project (Slovenia) with Riste Škrekovski, professor at the University of Ljubljana ended at the end of 2013. This project was entitled "Graphs for combinatorial chemistry and complex networks". Several manuscripts are under submission.

- 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.
- At present, Mario Valencia is the international coordinator of the MathAmSud project 13MATH-07 "Structural an algebraic problems on graph theory" (2013–2015). This project is funded by the following research institutes: CNRS in France, MinCyT in Argentina, CAPES in Brazil and CMM in Chile.

8.2. National Initiatives

8.2.1. ANR

8.2.1.1. HEREDIA

Participant: Jean-Sébastien Sereni [contact person].

HEREDIA (http://www.liafa.univ-paris-diderot.fr/~sereni/Heredia/) is an ANR JCJC ("Jeunes Chercheurs") focusing on hereditary properties of graphs, which provide a general perspective to study graph properties. Several important general theorems are known and the approach offers an elegant way of unifying notions and proof techniques. Further, hereditary classes of graphs play a central role in graph theory. Besides their theoretical appeal, they are also particularly relevant from an algorithmic point of view. With Jean-Sébastien Sereni, the HEREDIA project involves Pierre Charbit (LIAFA, Paris), Louis Esperet (G-SCOP, Grenoble) and Nicolas Trotignon (LIP, Lyon).

8.2.1.2. Hybride

Participants: Luis-Felipe Melo, Amedeo Napoli, Chedy Raïssi, My Thao Tang, Mohsen Sayed, Yannick Toussaint [contact person].

The Hybride research project (http://hybride.loria.fr/) aims at developing new methods and tools for supporting knowledge discovery from textual data by combining methods from Natural Language Processing (NLP) and Knowledge Discovery in Databases (KDD). A key idea is to design an interacting and convergent process where NLP methods are used for guiding text mining and KDD methods are used for analyzing textual documents. NLP methods are mainly based on text analysis, and extraction of general and temporal information. KDD methods are based on pattern mining, e.g. itemsets and sequences, formal concept analysis and variations, and graph mining. In this way, NLP methods applied to some texts locate "textual information" that can be used by KDD methods as constraints for focusing the mining of textual data. By contrast, KDD methods can extract itemsets or sequences that can 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).

8.2.1.3. ISTEX

Participants: Luis-Felipe Melo, Amedeo Napoli, Yannick Toussaint [contact person].

ISTEX is a so-called "Initiative d'excellence" managed by CNRS and DIST ("Direction de l'Information Scientifique et Technique"). ISTEX aims at giving to the research and teaching community an on-line access to scientific publications in all the domains. Thus ISTEX is in concern with a massive acquisition of documentation such as journals, proceedings, corpus, databases...ISTEX-R is one research project within ISTEX in which is involved the Orpailleur team, with two other partners, namely the ATILF laboratory and the INIST Institute (both in Nancy). ISTEX-R aims at developing a new generation of tools for querying full-text documentation, analyzing their content or extracting information and knowledge units. A platform is currently under development to provide robust NLP tools for text processing, as well as methods in text mining and domain conceptualization.

8.2.1.4. Kolflow

Participants: Jean Lieber [contact person], Alice Hermann, Amedeo Napoli, Emmanuel Nauer, My Thao Tang, Yannick Toussaint.

Kolflow (http://kolflow.univ-nantes.fr/) is a 3-year basic research project taking place from February 2011 to July 2014, funded by French National Agency for Research (ANR), program ANR CONTINT. The aim of the project is investigation on man-machine collaboration in continuous knowledge-construction flows.

Kolflow partners are GDD (LINA Nantes), Silex (LIRIS Lyon), Orpailleur (Inria NGE/LORIA), Score (Inria NGE/LORIA), and Wimmics (Inria Sophia Antipolis).

8.2.1.5. PEPSI: Polynomial Expansions of Protein Structures and Interactions

Participants: David Ritchie, Marie-Dominique Devignes, Malika Smaïl-Tabbone.

The PEPSI ("Polynomial Expansions of Protein Structures and Interactions") project is a collaboration with Sergei Grudinin at Inria Grenoble (project Nano-D) and Valentin Gordeliy at the Institut de Biologie Structurale (IBS) in Grenoble. This four-year project funded by the ANR "Modèles Numériques" program involves developing computational protein modeling and docking techniques and using them to help solve the structures of large molecular systems experimentally (http://pepsi.gforge.inria.fr).

8.2.1.6. Termith

Participants: Luis-Felipe Melo, Yannick Toussaint [contact person].

Termith (http://www.atilf.fr/ressources/termith/) is an ANR Project which involves the following laboratories: ATILF, LIDILEM, LINA, INIST, Inria Saclay and Inria Nancy Grand Est. It aims at indexing documents belonging to different domain of Humanities. Thus, the project focuses on extracting term candidates (information extraction) and on disambiguation.

In the Orpailleur team, we are mainly concerned by information extraction using Formal Concept Analysis techniques, but also itemset or sequence extraction. 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 term and its corresponding domain.

8.2.1.7. Trajcan: a study of patient care trajectories

Participants: Elias Egho, Nicolas Jay [contact person], Amedeo Napoli, Chedy Raïssi.

Since 30 years, many patient classification systems (PCS) have been developed. These systems aim at classifying care episodes into groups according to different patient characteristics. In France, the so-called "Programme de Médicalisation des Systèmes d'Information" (PMSI) is a national wide PCS in use in every hospital. It systematically collects data about millions of hospitalizations. Though it is used for funding purposes, it includes useful information for public health domains such as epidemiology or health care planning.

The objective of the Trajcan project is to represent and analyze "patient care trajectories" (patient suffering from cancer limited to breast, colon, rectum, and lung cancers) and the associated healthcares. The data are related to patients receiving hospital cares in the "Bourgogne" region and using data from the PMSI. Such an analysis involves various data, e.g. type of cancer, number of visits, type of stays, hospitalization services and therapies used, and demographic factors, i.e. age, gender, place of residence.

One thesis is currently carried out on this subject whose objective is to design a knowledge discovery system working on multidimensional and sequential data for characterizing Patient Care Trajectories (PCT) [52], [62]. This thesis combines knowledge discovery and knowledge representation methods for improving the definition of patient care trajectories as temporal objects (sequential data mining). The overall objective is to improve decision support and healthcare in detecting for example typical or exceptional trajectories for planning with precision healthcare for a given population.

In parallel, Formal Concept Analysis techniques were used in conjunction with regression tree analysis to produce semi-automated classification of PCTs in the field of breast cancer in France [27].

8.2.2. Other National Initiatives and Collaborations

8.2.2.1. PEPS Cryo-CA

Participant: David Ritchie [Inria Nancy].

Cryo-CA is a two-year PEPS project ("Projets exploratoires pluridisciplinaires") funded by CNRS, involving a collaboration with cryo-electron microscopy experimentalists at the IGBMC ("Institut de Génétique et de Biologie Moléculaire et Cellulaire") in Strasbourg. People involved in the project with David Ritchie are Sergei Grudinin (Inria Grenoble), Annick Dejaegere (IGBMC, Strasbourg), and Patrick Schultz (IGBMC Strasbourg). The aim of the project is to encourage collaborations between experimentalists and computer scientists in order to advance the state of the art of computational algorithms in structural biology.

8.2.2.2. Towards the discovery of new nonribosomal peptides and synthetases

We have initiated a collaboration with researchers from the LIFL and Université Lille Nord de France. We collaborated on the NRPS toolbox [109]. Data was cleaned and integrated from various public and specific analysis programs. The resulting database should facilitate the process of knowledge discovery of new nonribosomal peptides and synthetases.

8.3. Regional Initiatives

8.3.1. Le Bois Santé (LBS)

Participants: Emmanuel Bresso, Marie-Dominique Devignes [contact person], Malika Smaïl-Tabbone.

The project "LBS – Le Bois Santé – #38017" is funded by the European Regional Development Fund (FEDER) and the French "Fonds Unique Interministériel (FUI)" in the framework of the BioProLor consortium. This project is coordinated by "Harmonic Pharma", a start-up specialized in the identification of active principles in natural products. The aim of LBS is to exploit wood products in the pharmaceutical and nutriment domains. Concerned people in the team are working on data management and knowledge discovery about new therapeutic applications.

The BioProLor consortium is composed of 5 enterprises and 7 academic research teams, which were funded for 3 years (2010–2013) by AME ("Agence pour la Mobilisation Economique") for the design of compounds with high added-value which originate from plants in Lorraine. Finally, it should be noticed that the PhD Thesis work of Emmanuel Bresso was taken in charge by Harmonic Pharma (CIFRE contract, 2009-2013).

8.3.2. PEPS Mirabelle EXPLOD-Biomed

Participants: Adrien Coulet, Marie-Dominique Devignes [contact person], Gabin Personeni, Malika Smaïl-Tabbone.

This project initiates a collaboration with geneticists from the Hospital of Nancy, namely Philippe Jonveaux and Céline Bonnet. The aim of the EXPLOD-Biomed project is to propose novel knowledge discovery methods applied to Linked Open Data for discovering gene that could be responsible for intellectual deficiencies. Linked Open Data are available on-line, interconnected and encoded in a format which can be straightforwardly mapped to ontologies. Thus they offer novel opportunities for knowledge discovery in biomedical data. Here, geneticists are playing the role of experts, guiding the different steps of the knowledge discovery process.

8.3.3. Hydreos

Participant: Jean-François Mari [contact person].

The research project Hydreos (http://www.hydreos.fr/fr) is aimed at evaluating the quality of water. Actually, water resources relies on many agronomic variables, including land use successions. Accordingly, one objective of this research project is to have a better understanding of the changes in the organization of a territory. 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.

The systems ARPEnTAge (see § 5.2.2) and CarottAge (see § 5.2.1) are used in this context, especially by agronomists of INRA (ASTER Mirecourt http://www6.nancy.inra.fr/sad-aster and UMR Costel Rennes http://www.univ-rennes2.fr/costel). In addition, we participated in various meetings of researchers involved in the study of quality of groundwater in Alsace-Lorraine.

This year, or research work focused on collecting and preprocessing satellite data sampled in a territory in Brittany where there is an important phytoplanktonic biomass and Ulva species mass proliferation risk.

8.3.4. Contrat Plan État Région" (CPER)

The links between the Regional Administration and LORIA are materialized through the so-called "Contrat Plan État Région" (CPER) which is running from 2007 to 2013. The associated scientific program is called "Modélisations, informations et systèmes numériques" (MISN) and includes two tracks in which the Orpailleur team is involved.

- "Modeling Bio-molecules and their Interactions" (MBI).

  The general objective of this project is to study how domain knowledge can be taken into account for improving the modeling of biomolecules and their interactions, and the modeling of biological systems (http://bioinfo.loria.fr). Six scientific projects are currently under development and involve collaborations with computer scientists and people working either in biology or chemistry. This project is coordinated by Marie-Dominique Devignes.

- An Inria experimental research platform is currently developed in the framework of MBI (http://bioinfo.loria.fr/Plateforme%20MBI), which is aimed at sharing data and computing resources. The specific features of this platform are relative to biomolecules modeling, to classification and to data integration for data mining. The platform is a constituent of the North-East node of RENABI –"Réseau National des Plateformes Bioinformatiques"– together with the platforms in Strasbourg, Reims, Lille, and Nancy-INIST.

- "Traitement Automatique des Langues et des Connaissances" (TALC).

  TALC stands for "Automatic Processing of Languages and Knowledge". The general objective of TALC is to study the relations existing between knowledge discovery, knowledge representation, reasoning, and natural language processing. In the framework of TALC, the Orpailleur team plays an important role as the research themes of the team are closely related to those of TALC. Actually, research projects are currently under development on knowledge management and decision support involving in particular the Kasimir and the Taaable systems.

9. Dissemination

9.1. Scientific Animation

- The scientific animation in the Orpailleur team is based on two seminars, the Team Seminar and the BINGO seminar. The Team Seminar is held twice a month and is used either for general presentations of members of the team or for invited presentations of external researchers. The BINGO seminar is also held twice a month and is used for more specific presentations focusing on biological, chemical, and medical topics. Actually, both seminars are active and are useful instruments for researchers in the team.
- Members of the Orpailleur team are all involved, as members or as head persons, in various national research groups (mainly GDR CNRS I3 and BIM).
- The members of the Orpailleur team are involved in the organization of conferences and workshops, as members of conference program committees (ECAI, IJCAI, PKDD, ICFCA ...), as members of editorial boards, and finally in the organization of journal special issues.

- Valmi Dufour-Lussier, Emmanuelle Gaillard and Alice Hermann have organized the 21th French workshop on case-based reasoning in the framework of PFIA-2013 (*Plateforme IA 2013*). This workshop is an annual meeting gathering junior and senior researchers from different communities.
- Emmanuel Nauer co-organized the "Cooking with Computers" workshop at IJCAI 2013 (Beijing, China). This workshop aims at bringing together researchers from every possible fields of artificial intelligence applying their research on food and cooking.
- Amedeo Napoli organized with Sergei O. Kuznetsov (HSE Moscow) and Sebastian Rudolph (TU Dresden) the second workshop FCA4AI ("What can do FCA for Artificial Intelligence") which was associated with the last IJCAI Conference in Beijing (China, August 2013, http://ijcai13.org/program/workshop/30).
- Amedeo Napoli organized with Claudio Carpineto (Fondazione Ugo Bordoni, Roma) and Sergei O. Kuznetsov (HSE Moscow) a Workshop in Moscow, March 24 2013, on "Formal Concept Analysis meets Information Retrieval" (http://www.hse.ru/en/org/hse/fcair). This was a joint workshop with the ECIR Conference.
- Amedeo Napoli organized with with Francisco Carvalho of the organization of a Workshop in Recife (Brazil), April 24-26, namely the s"French-Brazilian Workshop on Numerical and Symbolic Methods of Data Analysis -WFB2013" (http://www.cin.ufpe.br/~wfb2013/). This workshop took place within an Inria-Facepe joint project where Amedeo Napoli and Francisco Carvalho are the scientific leaders.
- Amedeo Napoli was in charge of the organization of the so-called "Journées de la Société Francophone de ChemoInformatique" (SFCI) in October 2013 at LORIA (October 10th and 11th), which is the meeting of researchers working on chemoinformatics and computational chemistry in France, gathering around 100 persons (http://sfci2013.loria.fr/).
- Jean-Sébastien Sereni organized a workshop on combinatorics and distributed computing in the scope of the LEA STRUCO, with 21 participants, on November 12th–15h, 2013, in Pont-à-Mousson, France (http://www.liafa.univ-paris-diderot.fr/~sereni/STRUCO/CDCMeeting/).
- Mario Valencia has organized the Graph Days in Nancy on May 23th-24th 2013 (see http://www.loria.fr/news/journees-graphes-les-23-et-24-mai-2013-en).

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

- The members of the Orpailleur team are involved in teaching at all levels of teaching in University of Lorraine. Actually, most of the members of the Orpailleur team are employed on university 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, 11 p. [*DOI :* 10.1186/1471-2105-14-207], http://hal.inria.fr/hal-00843914

[2] A. COULET, F. DOMENACH, M. KAYTOUE, A. NAPOLI. *Using Pattern Structures for Analyzing Ontology-Based Annotations of Biomedical Data*, in "International Conference on Formal Concept Analysis", Dresden, Germany, Springer, May 2013, http://hal.inria.fr/hal-00880643

[3] 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), AAAI Press, August 2013, pp. 3002-3006, http://hal.inria.fr/hal-00871703

[4] Z. DVOŘÁK, J.-S. SERENI, J. VOLEC. , *Subcubic triangle-free graphs have fractional chromatic number at most* $14/5$, January 2013, http://hal.inria.fr/hal-00779634

[5] E. GAILLARD, J. LIEBER, Y. NAUDET, E. NAUER. *Case-Based Reasoning on E-Community Knowledge*, in "21st International Conference on Case-Based Reasoning", Saratoga Springs, NY, United States, Springer, 2013, vol. 7969, pp. 104-118 [*DOI :* 10.1007/978-3-642-39056-2_8], http://hal.inria.fr/hal-00918518

[6] A. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *KBDOCK 2013: A spatial classification of 3D protein domain family interactions*, in "Nucleic Acids Research", November 2013, http://hal.inria.fr/hal-00920612

[7] N. JAY, G. NUEMI, M. GADREAU, C. QUANTIN. *A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer*, in "BMC Medical Informatics and Decision Making", November 2013, vol. 13, n^{o} 1, 130 p. [*DOI :* 10.1186/1472-6947-13-130], http://hal.inria.fr/inserm-00917359

[8] C. LOW-KAM, C. RAÏSSI, M. KAYTOUE, J. PEI. *Mining Statistically Significant Sequential Patterns*, in "IEEE International Conference on Data Mining", Dallas, United States, December 2013, http://hal.inria.fr/hal-00922255

[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], http://hal.inria.fr/hal-00807178

[10] 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

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] E. BRESSO. , *Organisation et exploitation des connaissances sur les réseaux d'interactions biomoléculaires pour l'étude de l'étiologie des maladies génétiques et la caractérisation des effets secondaires de principes actifs*, Université de Lorraine, September 2013, http://hal.inria.fr/tel-00917934

[12] T. MEILENDER. , *Un wiki sémantique pour la gestion des connaissances décisionnelles - Application à la cancérologie*, Université de Lorraine, June 2013, http://hal.inria.fr/tel-00919997

Articles in International Peer-Reviewed Journals

[13] N. AGRINIER, C. ALTIERI, F. ALLA, N. JAY, D. DOBRE, N. THILLY, F. ZANNAD. *Effectiveness of a multidimensional home nurse led heart failure disease management program-A French nationwide time-series comparison*, in "International Journal of Cardiology", October 2013, vol. 168, n° 4, pp. 3652-8 [*DOI :* 10.1016/J.IJCARD.2013.05.090], http://hal.inria.fr/hal-00903263

[14] 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° 1, 11 p. [*DOI :* 10.1186/1471-2105-14-207], http://hal.inria.fr/hal-00843914

[15] A. CARRIERI, V. PÉREZ-NUENO, G. LENTINI, D. RITCHIE. *Recent trends and future prospects in computational GPCR drug discovery: from virtual screening to polypharmacology*, in "Current Topics in Medicinal Chemistry", 2013, vol. 13, n° 9, pp. 1069-1097 [*DOI :* 10.2174/15680266113139990028], http://hal.inria.fr/hal-00880351

[16] V. COEVOET, J. FRESSON, R. VIEUX, N. JAY. *Socioeconomic deprivation and hospital length of stay: a new approach using area-based socioeconomic indicators in multilevel models*, in "Medical Care", June 2013, vol. 51, n° 6, pp. 548-54 [*DOI :* 10.1097/MLR.0B013E3182928F84], http://hal.inria.fr/hal-00903258

[17] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Decentralized case-based reasoning and Semantic Web technologies applied to decision support in oncology*, in "Knowledge Engineering Review", March 2013, vol. 28, n° 4, pp. 425–449 [*DOI :* 10.1017/S0269888913000027], http://hal.inria.fr/hal-00922080

[18] H. DE ALMEIDA, I. BASTOS, B. RIBEIRO, B. MAIGRET, J. SANTANA. *New Binding Site Conformations of the Dengue Virus NS3 Protease Accessed by Molecular Dynamics Simulation*, in "PLoS ONE", August 2013, vol. 8, n° 8, http://hal.inria.fr/hal-00922971

[19] L. GHEMTIO, V. I. PÈREZ-NUENO, V. LEROUX, Y. ASSESS, M. SOUCHET, L. MAVRIDIS, B. MAIGRET, D. RITCHIE. *Recent trends and applications in 3D virtual screening*, in "Combinatorial Chemistry & High Throughput Screening", November 2013, vol. 15, n° 9, pp. 749-769, http://hal.inria.fr/hal-00923017

[20] A. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *KBDOCK 2013: A spatial classification of 3D protein domain family interactions*, in "Nucleic Acids Research", November 2013, http://hal.inria.fr/hal-00920612

[21] A. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Protein Docking Using Case-Based Reasoning*, in "Proteins", October 2013, vol. 81, n° 12, pp. 2150-2158 [*DOI :* 10.1002/PROT.24433], http://hal.inria.fr/hal-00880341

[22] S. GUESSOUM, M. T. LASKRI, J. LIEBER. *RespiDiag: a Case-Based Reasoning System for the Diagnosis of Chronic Obstructive Pulmonary Disease*, in "Expert Systems with Applications", February 2014, vol. 41, pp. 267–273, http://hal.inria.fr/hal-00912641

[23] T. V. HOANG, E. H. BARNEY SMITH, S. TABBONE. *Sparsity-based edge noise removal from bilevel graphical document images*, in "International Journal on Document Analysis and Recognition", August 2013, http://hal.inria.fr/hal-00852418

[24] T. V. HOANG, X. CAVIN, D. RITCHIE. *gEMfitter: A highly parallel FFT-based 3D density fitting tool with GPU texture memory acceleration*, in "Journal of Structural Biology", September 2013 [*DOI :* 10.1016/J.JSB.2013.09.010], http://hal.inria.fr/hal-00866871

[25] T. V. HOANG, X. CAVIN, P. SCHULTZ, D. RITCHIE. *gEMpicker: A Highly Parallel GPU-Accelerated Particle Picking Tool for Cryo-Electron Microscopy*, in "BMC Structural Biology", October 2013, http://hal.inria.fr/hal-00872625

[26] T. V. HOANG, S. TABBONE. *Errata and comments on "Generic orthogonal moments: Jacobi–Fourier moments for invariant image description"*, in "Pattern Recognition", April 2013, vol. 46, n° 11, pp. 3148-3155 [*DOI :* 10.1016/J.PATCOG.2013.04.011], http://hal.inria.fr/hal-00820279

[27] N. JAY, G. NUEMI, M. GADREAU, C. QUANTIN. *A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer*, in "BMC Medical Informatics and Decision Making", November 2013, vol. 13, n° 1, 130 p. [*DOI :* 10.1186/1472-6947-13-130], http://hal.inria.fr/inserm-00917359

[28] A.-S. KARABOGA, F. PETRONIN, G. MARCHETTI, M. SOUCHET, B. MAIGRET. *Benchmarking of HPCC: A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments*, in "Journal of Molecular Graphics and Modelling", January 2013, vol. 41, pp. 20-30, http://hal.inria.fr/hal-00922973

[29] M. LENSINK, I. MOAL, P. BATES, P. KASTRITIS, A. MELQUIOND, E. KARACA, C. SCHMITZ, M. VAN DIJK, A. BONVIN, M. EISENSTEIN, B. JIMÉNEZ-GARCÍ, S. GROSDIDIER, A. SOLERNOU, L. PÉREZ-CANO, C. PALLAR, J. FERNÁNDEZ-RECIO, J. XU, P. MUTHU, K. PRANEETH KILAMBI, J. GRAY, S. GRUDININ, G. DEREVYANKO, J. MITCHELL, J. WIETING, E. KANAMORI, Y. TSUCHIYA, Y. MURAKAMI, J. SARMIENTO, D. STANDLEY, M. SHIROTA, K. KINOSHITA, H. NAKAMURA, M. CHAVENT, D. RITCHIE, H. PARK, D. RITCHIE, J. KO, H. LEE, C. SEOK, Y. SHEN, D. KOZAKOV, S. VAJDA, P. KUNDROTAS, I. VAKSER, B. PIERCE, H. HWANG, T. VREVEN, Z. WENG, I. BUCH, E. FARKASH, H. WOLFSON, M. ZACHARIAS, S. QIN, H.-X. ZHOU, S.-Y. HUANG, X. ZOU, J. WOJDYLA, C. KLEANTHOUS, S. WODAK. *Blind prediction of interfacial water positions in CAPRI*, in "Proteins", October 2013 [*DOI :* 10.1002/PROT.24439], http://hal.inria.fr/hal-00880345

[30] 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], http://hal.inria.fr/hal-00807178

[31] P. POPOV, D. RITCHIE, S. GRUDININ. *DockTrina: Docking triangular protein trimers*, in "Proteins: Structure, Function, and Genetics", January 2014, vol. 82, n° 1, pp. 34-44 [*DOI :* 10.1002/PROT.24344], http://hal.inria.fr/hal-00880359

[32] J.-H. RAMAROSON, F. LE BER, B. RAMAMONJISOA, D. HERVÉ. *Treillis de Galois pour la fusion de connaissances spatiales sur des territoires villageois malgaches*, in "Revue d'Intelligence Artificielle", 2013, vol. 2013, n° 4-5, pp. 595-617, http://hal.inria.fr/hal-00862250

[33] 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° 1, pp. 81-108 [*DOI :* 10.1007/s10472-012-9329-3], http://hal.inria.fr/lirmm-00816300

[34] J.-S. SERENI, Z. YILMA. *A Tight Bound on the Set Chromatic Number*, in "Discussiones Mathematicae Graph Theory", 2013, vol. 33, n° 2, pp. 461–465 [*DOI :* 10.7151/DMGT.1679], http://hal.inria.fr/hal-00624323

[35] P. TORRES, M. VALENCIA-PABON. *On the packing chromatic number of hypercubes*, in "Electronic Notes in Discrete Mathematics", November 2013, vol. 44, n° 5, pp. 263-268, http://hal.inria.fr/hal-00926875

Articles in National Peer-Reviewed Journals

[36] F. KOHLER, N. JAY, F. DUCREAU, G. CASANOVA, C. KOHLER, A.-C. BENHAMOU. *COURLIS (COURs en LIgne de Statistiques appliquées) : un MOOC francophone innovant*, in "HEGEL", 2013, vol. 3, n° 1, pp. 27-32 [*DOI :* 10.4267/2042/49205], http://hal.inria.fr/hal-00903262

International Conferences with Proceedings

[37] M. ALAM, M. WUDAGE CHEKOL, A. COULET, A. NAPOLI, M. SMAÏL-TABBONE. *Lattice Based Data Access (LBDA): An Approach for Organizing and Accessing Linked Open Data in Biology*, in "DMoLD'13 - Data Mining on Linked Data Workshop (ECML/PKDD, 2013)", Prague, Czech Republic, Springer, September 2013, http://hal.inria.fr/hal-00903450

[38] J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Computing Similarity Dependencies with Pattern Structures*, in "The Tenth International Conference on Concept Lattices and their Applications - CLA 2013", La Rochelle, France, France, M. OJEDA-ACIEGO, J. OUTRATA (editors), CEUR Workshop Proceedings Vol. 1062, Karell Bertet, 2013, pp. 33-44, http://hal.inria.fr/hal-00922592

[39] S. BASU ROY, I. LYKOURENTZOU, S. THIRUMURUGANATHAN, S. AMER-YAHIA, G. DAS. *Crowds, not Drones: Modeling Human Factors in Interactive Crowdsourcing*, in "DBCrowd 2013 - VLDB Workshop on Databases and Crowdsourcing", Riva del Garda, Trento, Italy, R. CHENG, A. D. SARMA, S. MANIU, P. SENELLART (editors), CEUR Workshop Proceedings, CEUR-WS, August 2013, pp. 39-42, http://hal.inria.fr/hal-00923542

[40] A. BUZMAKOV, E. EGHO, N. JAY, S. O. KUZNETSOV, A. NAPOLI, C. RAÏSSI. *FCA and pattern structures for mining care trajectories*, in "Workshop FCA4AI, "What FCA can do for artificial intelligence?"", Beijing, China, August 2013, http://hal.inria.fr/hal-00910290

[41] A. BUZMAKOV, E. EGHO, N. JAY, S. O. KUZNETSOV, A. NAPOLI, C. RAÏSSI. *On Projections of Sequential Pattern Structures (with an application on care trajectories)*, in "The Tenth International Conference on Concept Lattices and Their Applications - CLA'13", La Rochelle, France, October 2013, http://hal.inria.fr/hal-00910300

[42] A. BUZMAKOV, E. EGHO, N. JAY, S. O. KUZNETSOV, A. NAPOLI, C. RAÏSSI. *The representation of sequential patterns and their projections within Formal Concept Analysis*, in "Workshop Notes for LML (PKDD)", Prague, Czech Republic, September 2013, http://hal.inria.fr/hal-00910266

[43] A. BUZMAKOV, A. NEZNANOV. *Practical Computing with Pattern Structures in FCART Environment*, in "Workshop FCA4AI, "What FCA can do for artificial intelligence?"", Beijing, China, August 2013, http://hal.inria.fr/hal-00910296

[44] M. W. CHEKOL, M. ALAM, A. NAPOLI. *A Study on the Correspondence between FCA and ELI Ontologies*, in "CLA - The Tenth International Conference on Concept Lattices and Their Applications - 2013", La Rochelle, France, October 2013, http://hal.inria.fr/hal-00906757

[45] M. W. CHEKOL, J. EUZENAT, P. GENEVÈS, N. LAYAÏDA. *Evaluating and benchmarking SPARQL query containment solvers*, in "Proc. 12th International semantic web conference (ISWC)", Sydney, Australia, H. ALANI, L. KAGAL, A. FOKOUE, P. GROTH, C. BIEMANN, J. X. PARREIRA, L. AROYO, N. NOY, C. WELTY, K. JANOWICZ (editors), Lecture notes in computer science, Springer Verlag, 2013, vol. 8219, pp. 408-423 [*DOI :* 10.1007/978-3-642-41338-4_26], http://hal.inria.fr/hal-00917911

[46] M. W. CHEKOL, A. NAPOLI. *A Study on the Correspondence between FCA and ELI Ontologies*, in "ISWC - The 12th International Semantic Web Conference - 2013", Sydney, Australia, October 2013, http://hal.inria.fr/hal-00906736

[47] M. W. CHEKOL, A. NAPOLI. *An FCA Framework for Knowledge Discovery in SPARQL Query Answers*, in "The 12th International Semantic Web Conference", Sydney, Australia, October 2013, http://hal.inria.fr/hal-00881080

[48] V. CODOCEDO, I. LYKOURENTZOU, H. ASTUDILLO, A. NAPOLI. *Using pattern structures to support information retrieval with Formal Concept Analysis*, in "International Workshop "What can FCA do for Artificial Intelligence?"", Beijing, China, S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors), August 2013, pp. 15-24, http://hal.inria.fr/hal-00880020

[49] A. COULET, F. DOMENACH, M. KAYTOUE, A. NAPOLI. *Using Pattern Structures for Analyzing Ontology-Based Annotations of Biomedical Data*, in "International Conference on Formal Concept Analysis", Dresden, Germany, Springer, May 2013, http://hal.inria.fr/hal-00880643

[50] M. D'AQUIN, N. JAY. *Interpreting data mining results with linked data for learning analytics: motivation, case study and directions*, in "LAK - Third International Conference on Learning Analytics and Knowledge - 2013", Leuven, Belgium, ACM, 2013, pp. 155–164 [*DOI :* 10.1145/2460296.2460327], http://hal.inria.fr/hal-00903259

[51] 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), AAAI Press, August 2013, pp. 3002-3006, http://hal.inria.fr/hal-00871703

[52] E. EGHO, N. JAY, C. RAÏSSI, G. NUEMI, C. QUANTIN, A. NAPOLI. *An approach for mining care trajectories for chronic diseases*, in "14th Conference on Artificial Intelligence in Medicine", Murcia, Spain, Springer, June 2013, vol. 7885, http://hal.inria.fr/hal-00883117

[53] E. GAILLARD, J. LIEBER, Y. NAUDET, E. NAUER. *Case-Based Reasoning on E-Community Knowledge*, in "21st International Conference on Case-Based Reasoning", Saratoga Springs, NY, United States, Springer, 2013, vol. 7969, pp. 104-118 [*DOI :* 10.1007/978-3-642-39056-2_8], http://hal.inria.fr/hal-00918518

[54] T. GUYET, F. LE BER, S. DA SILVA, C. LAVIGNE. *Comparaison des chemins de Hilbert adaptatif et des graphes de voisinage pour la caractérisation d'un parcellaire agricole*, in "Conférence Extraction et Gestion de Connaissances", Rennes, France, January 2014, http://hal.inria.fr/hal-00916964

[55] N. JAY, M. D'AQUIN. *Linked Data and Online Classifications to Organise Mined Patterns in Patient Data*, in "Proceedings of the AMIA 2013 Annual Symposium", Washington, United States, 2013, in press, http://hal.inria.fr/hal-00903261

[56] C. LOW-KAM, C. RAÏSSI, M. KAYTOUE, J. PEI. *Mining Statistically Significant Sequential Patterns*, in "IEEE International Conference on Data Mining", Dallas, United States, December 2013, http://hal.inria.fr/hal-00922255

[57] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *Soundness and Completeness of Relational Concept Analysis*, in "ICFCA'2013: 11th International Conference on Formal Concept Analysis", Dresden, Germany, P. CELLIER, F. DISTEL, B. GANTER (editors), Lecture Notes in Computer Science, Springer Netherlands, May 2013, vol. 7880, pp. 228-243 [*DOI :* 10.1007/978-3-642-38317-5_15], http://hal.inria.fr/lirmm-00833506

[58] R. STANLEY, H. ASTUDILLO, V. CODOCEDO, A. NAPOLI. *A Conceptual-KDD approach and its application to cultural heritage*, in "Concept Lattices and their Applications", La Rochelle, France, M. OJEDA-ACIEGO, J. OUTRATA (editors), L3i laboratory, University of La Rochelle, October 2013, pp. 163-174, http://hal.inria.fr/hal-00880002

[59] M. T. TANG, Y. TOUSSAINT. *A Collaborative Approach for FCA-Based Knowledge Extraction*, in "CLA - The Tenth International Conference on Concept Lattices and Their Applications - 2013", La Rochelle, France, October 2013, http://hal.inria.fr/hal-00906814

National Conferences with Proceedings

[60] J. COJAN, V. DUFOUR-LUSSIER, A. HERMANN, F. LE BER, J. LIEBER, E. NAUER, G. PERSONENI. *Révisor : un ensemble de moteurs d'adaptation de cas par révision des croyances*, in "JIAF - Septièmes Journées de l'Intelligence Artificielle Fondamentale - 2013", Aix-en-Provence, France, June 2013, http://hal.inria.fr/hal-00856487

[61] A. COULET, F. DOMENACH, M. KAYTOUE, A. NAPOLI. *Using pattern structures for analyzing ontology-based annotations of biomedical data*, in "Septièmes Journées d'Intelligence Artificielle Fondamentale", Aix-en-Provence, France, S. KONIECZNY, N. MAUDET, D. HABET, V. RISCH (editors), LSIS and LIF Marseille, June 2013, pp. 97–106, http://hal.inria.fr/hal-00922392

[62] E. EGHO, C. RAÏSSI, T. CALDERS, N. JAY, A. NAPOLI. *Vers une mesure de similarité pour les séquences complexes*, in "Extraction et gestion des connaissances (EGC'2013)", Toulouse, France, Cépaduès, February 2013, pp. 335-340, http://hal.inria.fr/hal-00885965

[63] R. HAFIANE, M. SMAÏL-TABBONE, M.-D. DEVIGNES, S. TABBONE. *Clustering optimal de gènes fondé sur une mesure de similarité sémantique*, in "10ème édition de la COnférence en Recherche d'Information et Applications - CORIA 2013", Neufchâtel, Switzerland, C. BERRUT (editor), 2013, 15 p. , http://hal.inria.fr/hal-00920700

[64] A. HERMANN, M. DUCASSÉ, S. FERRÉ, J. LIEBER. *Une approche fondée sur le raisonnement à partir de cas pour la mise à jour interactive d'objets du Web sémantique*, in "21ème atelier Français de Raisonnement à Partir de Cas (RàPC)", Lille, France, 2013, http://hal.inria.fr/hal-00910294

[65] G. PERSONENI, A. HERMANN, J. LIEBER. *Adaptation de cas propositionnels par réparations fondées sur des connaissances d'adaptation*, in "RàPC - 21ème atelier Français de Raisonnement à Partir de Cas - 2013", Lille, France, V. DUFOUR-LUSSIER, E. GAILLARD, A. HERMANN (editors), 2013, http://hal.inria.fr/hal-00910276

[66] M. T. TANG, Y. TOUSSAINT. *Vers un processus continu d'extraction de connaissances à partir de textes*, in "IC - The 24e journées francophones d'Ingénierie des Connaissances - 2013", Lille, France, July 2013, http://hal.inria.fr/hal-00861865

Conferences without Proceedings

[67] G. CANALS, A. CORDIER, E. DESMONTILS, L. INFANTE-BLANCO, E. NAUER. *Collaborative Knowledge Acquisition under Control of a Non-Regression Test System*, in "Workshop on Semantic Web Collaborative Spaces", Montpellier, France, May 2013, 14 p. , http://hal.inria.fr/hal-00880347

[68] E. GAILLARD, J. LIEBER, Y. NAUDET, E. NAUER. *Raisonner sur des connaissances provenant d'une e-communauté*, in "Ingénierie des connaissances 2013", Lille, France, 2013, http://hal.inria.fr/hal-00918540

Scientific Books (or Scientific Book chapters)

[69] 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], http://hal.inria.fr/hal-00912767

[70] D. RITCHIE, V. PÉREZ-NUENO. *ParaFit*, in "Scaffold Hopping in Medicinal Chemistry", N. BROWN (editor), Methods and Principles in Medicinal Chemistry, Wiley, 2013, vol. 58, http://hal.inria.fr/hal-00880352

Books or Proceedings Editing

[71] C. CARPINETO, S. O. KUZNETSOV, A. NAPOLI (editors). , *FCAIR 2012 Formal Concept Analysis Meets Information Retrieval Workshop co-located with the 35th European Conference on Information Retrieval (ECIR 2013) March 24, 2013, Moscow, Russia*, CEUR Proceedings, 2013, vol. 977, 157 p. , http://hal.inria.fr/hal-00922617

[72] S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors). , *International Workshop "What can FCA do for Artificial Intelligence?" (FCA4AI at IJCAI 2013, Beijing, China, August 4 2013)*, CEUR Proceedings, 2013, vol. 1058, 58 p. , http://hal.inria.fr/hal-00922616

Research Reports

[73] A. P. DOVE, J. R. GRIGGS, R. J. KANG, J.-S. SERENI. , *Supersaturation in the Boolean lattice*, March 2013, http://hal.inria.fr/hal-00802000

[74] Z. DVOŘÁK, J.-S. SERENI, J. VOLEC. , *Subcubic triangle-free graphs have fractional chromatic number at most* $14/5$, January 2013, http://hal.inria.fr/hal-00779634

[75] D. RAUTENBACH, J.-S. SERENI. , *Transversals of Longest Paths and Cycles*, February 2013, http://hal.inria.fr/hal-00793271

[76] J.-S. SERENI, J. VOLEC. , *A note on acyclic vertex-colorings*, December 2013, http://hal.inria.fr/hal-00921122

Other Publications

[77] F. BONOMO, O. SCHAUDT, M. STEIN, M. VALENCIA-PABON. , *b-coloring is NP-hard on co-bipartite graphs and polytime solvable on tree-cographs*, 2013, http://hal.inria.fr/hal-00926924

[78] V. DUFOUR-LUSSIER, B. GUILLAUME, G. PERRIER. , *Parsing Coordination Extragrammatically*, 2013, accepted for publication in LNAI (Lecture Notes in Artificial Intelligence), http://hal.inria.fr/hal-00921033

References in notes

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

[80] P. BUITELAAR, P. CIMIANO, B. MAGNINI (editors). , *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence and Applications, IOS PressAmsterdam, 2005

[81] S. STAAB, R. STUDER (editors). , *Handbook on Ontologies (Second Edition)*, SpringerBerlin, 2009

[82] M. BARBUT, B. MONJARDET. , *Ordre et classification – Algèbre et combinatoire (2 tomes)*, HachetteParis, 1970

[83] S. BENABDERRAHMANE, M. SMAÏL-TABBONE, O. POCH, A. NAPOLI, M.-D. DEVIGNES. *IntelliGO: a new vector-based semantic similarity measure including annotation origin*, in "BMC Bioinformatics", December 2010, vol. 11, n^{o} 1, 588 p. [*DOI :* 10.1186/1471-2105-11-588], http://www.biomedcentral.com/1471-2105/11/588/abstract, http://hal.inria.fr/inria-00543910/en

[84] R. BENDAOUD, A. NAPOLI, Y. TOUSSAINT. *A proposal for an Interactive Ontology Design Process based on Formal Concept Analysis*, in "Formal Ontology in Information Systems – Proceedings of the Fifth International Conference (FOIS 2008)", Amsterdam, C. ESCHENBACH, M. GRÜNINGER (editors), Frontiers in Artificial Intelligence and Applications, IOS Press, 2008, pp. 311–323

[85] R. BENDAOUD, A. NAPOLI, Y. TOUSSAINT. *Formal Concept Analysis: A unified framework for building and refining ontologies*, in "Knowledge Engineering: Practice and Patterns - Proceedings of the 16th International Conference EKAW", A. GANGEMI, J. EUZENAT (editors), Lecture Notes in Computer Science 5268, 2008, pp. 156–171

[86] R. BENDAOUD, Y. TOUSSAINT, A. NAPOLI. *PACTOLE: A methodology and a system for semi-automatically enriching an ontology from a collection of texts*, in "Proceedings of the 16th International Conference on Conceptual Structures, ICCS 2008, Toulouse, France", P. EKLUND, O. HAEMMERLÉ (editors), Lecture Notes in Computer Science 5113, 2008, pp. 203–216

[87] H. M. BERMAN, T. BATTISTUZ, T. N. BHAT, W. F. BLUHM, P. E. BOURNE, K. BURKHARDT, L. IYPE, S. JAIN, P. FAGAN, J. MARVIN, D. PADILLA, V. RAVICHANDRAN, B. SCHNEIDER, N. THANKI, H. WEISSIG, J. D. WESTBROOK, C. ZARDECKI. *The Protein Data Bank*, in "Acta Crystallographica Section D-Biological Crystallography", 2002, vol. 58, pp. 899–907

[88] P. CIMIANO, A. HOTHO, S. STAAB. *Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis*, in "Journal of Artificial Intelligence Research", 2005, vol. 24, pp. 305–339

[89] L. DE RAEDT. , *Logical and Relational Learning*, Cognitive Technologies, Springer, 2008

[90] M.-D. DEVIGNES, S. BENABDERRAHMANE, M. SMAÏL-TABBONE, A. NAPOLI, O. POCH. *Functional classification of genes using semantic distance and fuzzy clustering approach: Evaluation with reference sets and overlap analysis*, in "international Journal of Computational Biology and Drug Design. Special Issue on: "Systems Biology Approaches in Biological and Biomedical Research"", 2012, vol. 5, n^{o} 3/4, pp. 245-260, http://hal.inria.fr/hal-00734329

[91] R. D. FINN, J. MISTRY, J. TATE, P. COGGILL, A. HEGER, J. E. POLLINGTON, O. L. GAVIN, P. GUNASEKARAN, G. CERIC, K. FORSLUND, L. HOLM, E. L. L. SONNHAMMER, S. R. EDDY, A. BATEMAN. *The Pfam protein families database*, in "Nucleic Acids Research", 2010, vol. 38

[92] B. GANTER, S. O. KUZNETSOV. *Pattern Structures and Their Projections*, in "Conceptual Structures: Broadening the Base, Proceedings of the 9th International Conference on Conceptual Structures, ICCS 2001, Stanford, CA", H. DELUGACH, G. STUMME (editors), Lecture Notes in Computer Science 2120, Springer, 2001, pp. 129–142

[93] B. GANTER, R. WILLE. , *Formal Concept Analysis*, SpringerBerlin, 1999

[94] A. GHOORAH, M.-D. DEVIGNES, M. SMAÏL-TABBONE, D. RITCHIE. *Spatial clustering of protein binding sites for template based protein docking*, in "Bioinformatics", August 2011, vol. 27, n^{o} 20, pp. 2820-2827 [*DOI :* 10.1093/BIOINFORMATICS/BTR493], http://hal.inria.fr/inria-00617921

[95] A. GHOORAH. , *Extraction de Connaissances pour la Modelisation tri-dimensionnelle de l'Interactome Structural*, Université de Lorraine, November 2012, http://hal.inria.fr/tel-00762444

[96] P. HITZLER, M. KRÖTSCH, S. RUDOLPH. , *Foundations of Semantic Web Technologies*, CRC PressBoca raton (FL), 2009

[97] H. HWANG, T. VREVEN, J. JANIN, Z. WENG. *Protein-protein docking benchmark version 4.0*, in "Proteins: Structure Function and Bioinformatics", 2010, vol. 78, n^{o} 15, pp. 3111–3114

[98] C. JONQUET, P. LEPENDU, S. FALCONER, A. COULET, N. NOY, M. A. MUSEN, N. H. SHAH. *NCBO Resource Index: Ontology-Based Search and Mining of Biomedical Resources*, in "Journal of Journal of Web Semantics", Sep 2011, vol. 9, n^{o} 3, pp. 316–324, NIH Projet NCBO [*DOI :* 10.1016/J.WEBSEM.2011.06.005], http://hal-lirmm.ccsd.cnrs.fr/lirmm-00622155

[99] M. KAYTOUE, S. O. KUZNETSOV, J. MACKO, W. MEIRA, A. NAPOLI. *Mining Biclusters of Similar Values with Triadic Concept Analysis*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, A. NAPOLI, V. VYCHODIL (editors), Inria Nancy Grand Est - LORIA, 2011, http://hal.inria.fr/hal-00640873/en

[100] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Revisiting Numerical Pattern Mining with Formal Concept Analysis*, in "Twenty second International Joint Conference on Artificial Intelligence - IJCAI 2011", Barcelona, Spain, 2011, http://hal.inria.fr/inria-00584371/en

[101] M. KAYTOUE, F. MARCUOLA, A. NAPOLI, L. SZATHMARY, J. VILLERD. *The Coron System*, in "8th International Conference on Formal Concept Analsis (ICFCA) - Supplementary Proceedings", L. BOUMEDJOUT, P. VALTCHEV, L. KWUIDA, B. SERTKAYA (editors), 2010, pp. 55–58

[102] T. LEVIANDIER, A. ALBER, F. LE BER, H. PIÉGAY. *Comparison of statistical algorithms for detecting homogeneous river reaches along a longitudinal continuum*, in "Geomorphology", 2012, vol. 138, n^o 1, pp. 130-144 [*DOI :* 10.1016/J.GEOMORPH.2011.08.031], http://hal.inria.fr/hal-00640698

[103] J. LIEBER, M. D'AQUIN, F. BADRA, A. NAPOLI. *Modeling adaptation of breast cancer treatment decision protocols in the KASIMIR project*, in "Applied Intelligence", 2008, vol. 28, n^o 3, pp. 261–274

[104] J. LIEBER, A. NAPOLI, L. SZATHMARY, Y. TOUSSAINT. *First Elements on Knowledge Discovery guided by Domain Knowledge (KDDK)*, in "Concept Lattices and Their Applications (CLA 06)", S. B. YAHIA, E. M. NGUIFO, R. BELOHLAVEK (editors), Lecture Notes in Artificial Intelligence 4923, Springer, Berlin, 2008, pp. 22–41

[105] Y. LIU, A. COULET, P. LEPENDU, N. H. SHAH. *Using ontology-based annotation to profile disease research*, in "Journal of the American Medical Informatics Association", June 2012, vol. 19, n^o e1 [*DOI :* 10.1136/AMIAJNL-2011-000631], http://hal.inria.fr/hal-00752101

[106] G. MACINDOE, L. MAVRIDIS, V. VENKATRAMAN, M.-D. DEVIGNES, D. RITCHIE. *HexServer: an FFT-based protein docking server powered by graphics processors*, in "Nucleic Acids Research", May 2010, vol. 38, W445 p. [*DOI :* 10.1093/NAR/GKQ311], http://hal.inria.fr/inria-00522712/en

[107] 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

[108] A. NAPOLI. *A smooth introduction to symbolic methods for knowledge discovery*, in "Handbook of Categorization in Cognitive Science", H. COHEN, C. LEFEBVRE (editors), Elsevier, Amsterdam, 2005, pp. 913–933

[109] M. PUPIN, M. SMAÏL-TABBONE, P. JACQUES, M.-D. DEVIGNES, V. LECLÈRE. *NRPS toolbox for the discovery of new nonribosomal peptides and synthetases*, in "Journées Ouvertes en Biologie, l'Informatique et les Mathématiques - JOBIM 2012", Rennes, France, F. COSTE, D. TAGU (editors), 2012, pp. 89-93, http://hal.inria.fr/hal-00734312

[110] V. PÉREZ-NUENO, D. RITCHIE, J. BORRELL, J. TEIXIDÓ. *Clustering and Classifying Diverse HIV Entry Inhibitors Using a Novel Consensus Shape-Based Virtual Screening Approach: Further Evidence for Multiple Binding Sites within the CCR5 Extracellular Pocket*, in "Journal of Chemical Information and Modeling", 2008, vol. 48, n^o 11, pp. 2146–2165

[111] C. RAÏSSI, J. PEI, T. KISTER. *Computing Closed Skycubes*, in "Proceedings of the VLDB Endowment", September 2010, vol. 3, n^o 1, pp. 838–847, http://hal.inria.fr/inria-00610923/en

[112] D. RITCHIE, A. GHOORAH, L. MAVRIDIS, V. VENKATRAMAN. *Fast Protein Structure Alignment using Gaussian Overlap Scoring of Backbone Peptide Fragment Similarity*, in "Bioinformatics", October 2012, vol. 28, n^o 24, pp. 3274-3281 [*DOI :* 10.1093/BIOINFORMATICS/BTS618], http://hal.inria.fr/hal-00756813

[113] D. RITCHIE, G. KEMP. *Protein Docking Using Spherical Polar Fourier Correlations*, in "Proteins: Structure, Function and Genetics", 2000, vol. 39, n^o 2, pp. 178–194

[114] D. RITCHIE, V. VENKATRAMAN. *Ultra-fast FFT protein docking on graphics processors*, in "Bioinformatics", 2010, vol. 26, n^{o} 19, pp. 2398–2405 [*DOI :* 10.1093/BIOINFORMATICS/BTQ444], http://hal.inria.fr/inria-00537988/en/

[115] V. ROTH, J. LAUB, M. KAWANABE, J. M. BUHMANN. *Optimal cluster preserving embedding of nonmetric proximity data*, in "IEEE Trans. Pattern Analysis and Machine Intelligence", 2003, vol. 25

[116] N. SCHALLER, E.-G. LAZRAK, P. MARTIN, J.-F. MARI, C. AUBRY, M. BENOÎT. *Combining farmers' decision rules and landscape stochastic regularities for landscape modelling*, in "Landscape Ecology", March 2012, vol. 27, n^{o} 3, pp. 433-446 [*DOI :* 10.1007/S10980-011-9691-2], http://hal.inria.fr/hal-00656407

[117] L. SZATHMARY. , *Symbolic Data Mining Methods with the Coron Platform*, Université Henri Poincaré (Nancy 1), 2006

[118] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *Constructing Iceberg Lattices from Frequent Closures Using Generators*, in "Discovery Science", J.-F. BOULICAUT, M. BERTHOD, T. HORVÁTH (editors), Lecture Notes in Computer Science 5255, Springer, Berlin, 2008, pp. 136–147

[119] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *Efficient Vertical Mining of Frequent Closures and Generators*, in "Proceedings of the 8th International Symposium on Intelligent Data Analysis (IDA-2009), Lyon, France", N. ADAMS, J.-F. BOULICAUT, C. ROBARDET, A. SIEBES (editors), Lecture Notes in Computer Science 5772, Springer, Berlin, 2009, pp. 393–404

[120] L. SZATHMARY, P. VALTCHEV, A. NAPOLI. *Finding Minimal Rare Itemsets and Rare Association Rules*, in "Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management (KSEM-2010), Belfast, Northern Ireland, UK", Y. BI, M.-A. WILLIAMS (editors), Lecture Notes in Artificial Intelligence 6291, Springer, Berlin, 2010, pp. 16–27

[121] L. SZATHMARY, P. VALTCHEV, A. NAPOLI. *Generating Rare Association Rules Using the Minimal Rare Itemsets Family*, in "International Journal of Software and Informatics", 2010, vol. 4, n^{o} 3, pp. 219–238

[122] V. VENKATRAMAN, D. RITCHIE. *Flexible protein docking refinement using pose-dependent normal mode analysis*, in "Proteins", June 2012, vol. 80, n^{o} 9, pp. 2262-2274 [*DOI :* 10.1002/PROT.24115], http://hal.inria.fr/hal-00756809

[123] V. VENKATRAMAN, D. RITCHIE. *Predicting Multi-component Protein Assemblies Using an Ant Colony Approach*, in "International Journal of Swarm Intelligence Research", September 2012, vol. 3, pp. 19-31 [*DOI :* 10.4018/JSIR.2012070102], http://hal.inria.fr/hal-00756807

[124] Y. ZHANG, J. SKOLNICK. *TM-align: a protein structure alignment algorithm based on TM-score*, in "Nucleic Acids Research", 2005, vol. 33, n^{o} 7, pp. 2302–2309

[125] U. VON LUXBURG. *A tutorial on spectral clustering*, in "Statistics and Computing", 2007, vol. 17, n^{o} 4, pp. 395-416