IN PARTNERSHIP WITH:
CNRS

Université de Lorraine

Activity Report 2012

Project-Team ORPAILLEUR

Knowledge Discovery guided by Domain Knowledge

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, Senior Researcher, CNRS, HdR]
Marie-Dominique Devignes [Junior Researcher, CNRS, HdR]
Bernard Maigret [Senior Researcher (Emeritus), CNRS, HdR]
Chedy Raïssi [Junior Researcher, Inria]
Dave Ritchie [Senior Researcher, Inria, HdR]
Yannick Toussaint [Junior Researcher, Inria, HdR]

Faculty Members

Adrien Coulet [Associate Professor (Telecom Nancy, Université de Lorraine)]
Nicolas Jay [Associate Professor (Faculté de Médecine, Université de Lorraine)]
Florence Le Ber [Professor (ENGEES Strasbourg), HdR]
Jean Lieber [Associate Professor (Université de Lorraine), HdR]
Jean-François Mari [Professor (Université de Lorraine), HdR]
Emmanuel Nauer [Associate Professor (Université de Lorraine)]
Malika Smaïl-Tabbone [Associate Professor (Université de Lorraine)]

Engineers

Jérémie Bourseau [Engineer (ADT, since November 2012)]
Renaud Grisoni [Engineer (until September 2012)]
Laura Infante Blanco [Engineer (ADT)]
Jean-François Kneib [Engineer (until September 2012)]
Luis Felipe Melo [Engineer (ADT until November 2012, Hybride project)]

PhD Students

Mehwish Alam [PhD Student (BioIntelligence Grant)]
Aleksey Buzmakov [PhD Student (BioIntelligence Grant)]
Emmanuel Bresso [PhD Student (Cifre Harmonic Pharma)]
Victor Codocedo [PhD Student (Quaero Grant)]
Sébastien Da Silva [PhD Student (INRA - Inria Grant)]
Valmi Dufour-Lussier [PhD Student (MERT Grant)]
Elias Egho [PhD Student (ANR Trajcan Project Grant)]
Emmanuelle Gaillard [Engineer (from January to June 2012), PhD Student (MERT Grant since October 2012)]
Thomas Meilender [PhD Student (CIFRE Grant, A2ZI Company), ATER (Université de Lorraine since September 2012)]
Julien Stévenot [PhD Student (ANR Kolflow Grant, until August 2012)]
My Thao Tang [PhD Student (ANR Kolflow Grant)]
Yasmine Assess [Engineer (BioIntelligence Project until June 2012), ATER (Université de Lorraine, since November 2012)]
Anisah Ghoorah [PhD Student (ANR Contract until October 2012, ATER since September 2012, Thesis defended in November 2012)]

Post-Doctoral Fellows

Thomas Bourquard [BioIntelligence Project (until June 2012)]

Van-Thai Hoang [ANR PEPSI (since January 2012)]
Ioanna Lykourentzou [ERCIM Postdoc (until April 2012), associate researcher (since May 2012)]
Violeta Pérez-Nueno [ERC Marie Curie Fellow (until July 2012)]

Administrative Assistant
Emmanuelle Deschamps [Secretary]

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.

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 making a step towards the notion of *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 and reifying the notion of KDDK.

2.2. Highlights of the Year

A best paper award was granted to a paper published in the proceedings of ICCBR-2012 (the international conference on case-based reasoning) . This paper presents an approach for adapting cases in the formalism of qualitative algebras, with an application in a temporal algebra, dedicated to adaptation of cooking recipe preparations, and an application in a spatial algebra, dedicated to the allocation of crops in a farmland.
BEST PAPER AWARD :

3. Research Program

3.1. From KDD to KDDK

knowledge discovery in databases, knowledge discovery in databases guided by domain knowledge, data mining methods

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.

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, and concept lattice design (Formal Concept Analysis and extensions [95], [108]).
- Numerical methods are based on second-order Hidden Markov Models (HMM2, designed for pattern recognition [107]). Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data.

The principle summarizing KDDK can be understood as a process going from complex data units to knowledge units being guided by domain knowledge (KDDK or "knowledge with/for knowledge") [104]. Two original aspects can be underlined: (i) the KDD process is guided by domain knowledge, and (ii) the extracted units are embedded within a knowledge representation formalism to be reused in a knowledge-based system for problem solving purposes.

The various instantiations of the KDDK process in the research work of Orpailleur are mainly based on *classification*, considered as a polymorphic process involved in tasks such as modeling, mining, representing, and reasoning. Accordingly, the KDDK process may feed knowledge-based systems to be used for problem-solving activities in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine, and also for semantic web activities involving text mining, information retrieval, and ontology engineering [97], [81].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

knowledge discovery in databases guided by domain knowledge, lattice-based classification, formal concept analysis, frequent itemset search, association rule extraction, second-order Hidden Markov Models, stochastic process, numerical data mining method

knowledge discovery in databases guided by domain knowledge is a KDD process guided by domain knowledge ; the extracted units are represented within a knowledge representation formalism and embedded within a knowledge-based system.

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 (from a binary database) a set of formal concepts organized within a concept lattice [95]. Concept lattices are sometimes 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 [123], [122].

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 [124], [125].

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. These 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. A special research effort focuses on the combination of knowledge elicited by experts and time-space regularities as extracted by an unsupervised classification based on stochastic models [23].

3.3. Elements on Text Mining

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

Text mining is a process for extracting knowledge units from large collections of texts, units that can be interpreted and reused within knowledge-based systems.

The objective of a text mining process is to extract new and useful knowledge units in a large set of texts [80], [89]. 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 the mining process more complex. 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, the text mining process is aimed at extracting new knowledge units from texts with the help of background knowledge encoded within an ontology and which is useful to relate notions present in a text, to guide and to help the text mining process. 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 real-world texts in application domains such as astronomy, 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

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

Knowledge representation is a process for representing knowledge within an ontology using a knowledge representation formalism, giving knowledge units a syntax and a semantics. Semantic web is based on ontologies and allows search, manipulation, and dissemination of documents on the web by taking into account their contents, i.e. the semantics of the elements included in the documents.

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 very difficult and tedious. Semantic web is an attempt for guiding search for information with the help of machines, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. However, a machine 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, i.e. describing the content of documents, and giving a semantics to this content. Knowledge representation languages are (the?) 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 (or DL [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 that is 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 associated to case-based reasoning (CBR), that 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.

In the trend of semantic web, research work is also carried on semantic wikis which are wikis i.e., 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.

4. Application Domains

4.1. Life Sciences

Participants: Yasmine Assess, Thomas Bourquard, Emmanuel Bresso, Marie-Dominique Devignes, Elias Egho, Anisah Ghoorah, Renaud Grisoni, Nicolas Jay, Bernard Maigret, Amedeo Napoli, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Yannick Toussaint.

knowledge discovery in life sciences, bioinformatics, biology, chemistry, gene

Knowledge discovery in life sciences is a process for extracting knowledge units from large biological databases, e.g. collection of genes.

One major application domain which is currently investigated by 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 and thus KDDK. 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. Solving problems for biologists using KDDK methods involves the design of specific modules that, in turn, leads to adaptations of the KDDK process, especially in the preparation of data and in the interpretation of the extracted units.

4.2. Knowledge Management in Medicine

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

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 ("Centre 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 have been used and adapted in the Kasimir project for several years. Recently, a semantic wiki has been deployed in its production version (http://www.oncologik.fr) It allows the management of decision protocols [48] More precisely, the migration from the static HTML site of Oncolor to a semantic wiki (with limited editing rights and unlimited reading rights) has been done [49]. This has consequences on the editorial chain of the published protocols which is more collaborative. A decision tree editor that has been integrated into the wiki and that has an export facility to formalized protocols in OWL DL has also been developed [61].

4.3. Cooking

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

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

The origin of the Taaable project is the Computer Cooking Contest (CCC). A contestant of the CCC is a system that answers queries of 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 that uses various technologies used and developed in the Orpailleur team, such as technologies of the semantic web, knowledge discovery techniques, knowledge representation and reasoning techniques, etc. From a research viewpoint it enables to test the scientific results on an application domain that is at the same time simple to understand and raising complex issues, and to study the complementarity of various research domains. Taaable has been at the origin of the project Kolflow of the ANR CONTINT program, whose application domain is WikiTaaable, the semantic wiki of Taaable. It is also used for other projects under submission.

5. Software and Platforms

5.1. Generic Symbolic KDD Systems

5.1.1. The Coron Platform

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

data mining, frequent itemsets, frequent closed itemsets, frequent generators, association rule extraction, rare itemsets

The Coron platform [120], [102] 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, frequent closed itemsets, frequent generators. In this collection we can find APriori, Close, Pascal, Eclat, Charm, and, as well, original algorithms such as ZART, Snow, Touch, and Talky-G. The Coron-base component contains also algorithms for extracting rare itemsets and rare association rules, e.g. APriori-rare, MRG-EXP, ARIMA, and BTB. 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].

skyline, skycube algorithms

This program implements the algorithms described in a research paper published last year at VLDB 2010 [112]. 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].

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

It provides 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. A new version incorporating a graphic user interface was released and is now running on Windows systems.

CarottAge has been 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 [17].

5.2.2. The ARPEnTAge system

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

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 [106]. 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.

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

We model the spatial structure of the landscape by a Potts model with external field whose sites are LUS located in the parcels. The dynamics of these LUS are modeled by a temporal HMM2. This leads to the definition of a Potts model where the underlying mean field is approximated by a hierarchical hidden Markov model that processes a Hilbert-Peano fractal curve spanning the image.

Those stochastic models have been used to segment the landscape into patches, each of them being characterized by a temporal HMM2. The patch labels, together with the geographic coordinates, determine a clustered image of the landscape that can be coded within an ESRI shapefile. ARPEnTAge can locate in a 2-D territory time regularities and implements a Time-dominant approach in Geographic Information Systems.

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 have been used for mining decision rules in a territory holding environmental issues. They provide a way of visualizing the impact of farmers decision rules in the landscape and revealing new extra hidden decision rules [23].

5.2.3. GenExp-LandSiTes: KDD and simulation

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

simulation, Hidden Markov Models

In the framework of the project “Impact des OGM” initiated by the French ministry of research, we have developed a software called GenExp-LandSiTes for simulating bidimensional random landscapes, and then studying the dissemination of vegetable transgenes. The GenExp-LandSiTes system is linked to the CarottAge system, and is based on computational geometry and spatial statistics. The simulated landscapes are given as input for programs such as “Mapod-Maïs” or “GeneSys-Colza” for studying the transgene diffusion. Other landscape models based on tessellation methods are under studies. The last version of GenExp allows an interaction with R and deals with several geographical data formats.

This work is now part of an INRA research network about landscape modeling, PAYOTE, that gathers several research teams of agronomists, ecologists, statisticians, and computer scientists. Sébastien da Silva is preparing his PhD thesis within this framework and is conducted both by Claire Lavigne (DR in ecology, INRA Avignon) and Florence Le Ber [46], [40].

GenExp-LandSiTes was part of a survey about innovative tools for geographical information [74], [73]. This survey has been conducted within the GDR Magis and has been presented in a book both in French and in English.

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 made available on line (http://plateforme-mbi.loria.fr/intelligo/) to be used by members of the community for exploitation and 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 has been developed using rigorous software engineering practices and is used by professionals in both industry and academia. The KNIME environment 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.

A position of engineer ("Ingénieur Jeune Diplomé Inria") was granted to the Orpailleur team to develop some extra KNIME nodes for relational data mining using the ALEPH program (http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/aleph.pl). The developed KNIME 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. An Inria APP procedure is currently pending.

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

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

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 steady-version of the software has been deposited 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. A poster was presented at the last JOBIM conference (Paris, June 2011). Recently, MODIM was used by colleagues from the LIFL for organizing data about non ribosomal peptide syntheses. Feedback from users led to extensions of the software. 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.

classification-based reasoning, case-based reasoning, edition and maintenance of knowledge, decision knowledge management, 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 of 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) has been 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.

The software 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 implemented thanks to a frequent close itemset extraction module of the Coron platform (see §5.1.1).

The Oncologik system is a collaborative editing tool aiming at facilitating the management of medical guidelines [49], [48]. Based on a semantic wiki, it allows the acquisition of formalized decision knowledge. A production version was released this year (http://www.oncologik.fr/). Oncologik also includes a graphical decision tree editor, KcatoS [61].

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].

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

Taaable 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 of knowledge engineering, namely 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, have been 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 since 2008 on line at http://taaable.fr, but is constantly evolving. This year, Taaable has been extended by two new features, both concerning knowledge acquisition.

The first feature uses closed itemsets for extracting adaptation knowledge in order to better adapt recipes. A first approach integrates a previous work about adaptation rule extraction [93] into a collaborative environment, in which humans and machines may now collaborate to better acquire adaptation rules [38]. This environment integrates also the results of a new work on knowledge extraction where specific cooking adaptation rules that can be applied to a single recipe, are generalized using close itemsets into generic adaptation rules, to make them usable on other recipes [60].

The second feature addresses the improvement of the formal representation of the preparation part of recipes, using a semi-automatic annotation process [59]. In Taaable, the procedural text describing the preparation is formalized in a graph, where cooking actions and ingredients, among others, are represented as vertexes, and semantic relations between those, shown as arcs. As the automatic annotation process that transforms, using natural language processing, a procedural text into a graph, produces incomplete annotation (disconnected graphs) or other annotation errors, a validating and correcting step is required. A specific graphical interface has been built to provide the users with a way to correct the graph representation of the cooking process, improving at the same time the quality of the knowledge about cooking procedures.

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

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

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, 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 that will ensure facilities for being plugged over data available on the web. The domain knowledge is considered to be represented in a RDF store, which could be additionally be interfaced with a semantic wiki, in order to benefit from the collaborative edition and management of the knowledge involved in the reasoning system (cases, ontology, adaptation rules). This development is support by an Inria ADT funding.

6. New Results

6.1. The Mining of Complex Data

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

formal concept analysis, relational concept analysis, pattern structures, search for frequent itemsets, association rule extraction, mining of complex data, graph mining, skylines, sequence mining, FCA in spatial and temporal reasoning

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 may be 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 working on extensions of such symbolic data mining methods to be applied on complex data such as 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, RCA, and Pattern Structures

Recent advances in data and knowledge engineering have emphasized the need for Formal Concept Analysis (FCA) tools taking into account structured data. 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 [116]. 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) [94], which allows to build a concept lattice from complex data, e.g. nominal, numerical, and interval data. In [101]), pattern structures are used for building a concept lattice from intervals, in full compliance with FCA, thus benefiting of the efficiency of FCA algorithms. Actually, the notion of similarity between objects is closely related to these extensions of FCA: 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). Various results were obtained in the study of the relations existing between FCA with an embedded explicit similarity measure and FCA with pattern structures [100]. Moreover, similarity is not a transitive relation and this lead us to the study of tolerance relations. In addition, a new research perspective is aimed at using frequent itemset search methods for mining interval-based data being guided by pattern structures and biclustering as well.

6.1.2. Advances in FCA and Pattern Mining

In the context of environmental sciences, research work is in concern with the mining of complex hydroecological data with concept lattices. In particular, Florence Le Ber –as a member of UMR 7517 Lhyges, Strasbourg– is the scientific head of an ANR project named "FRESQUEAU" (2011–2014) dealing with FCA and data mining and hydroecological data (see http://engees.unistra.fr/site/recherche/projets/anr-fresqueau/).

In this framework, concept lattices based on multi-valued contexts have been used for characterizing macroinvertebrate communities in wetland and their seasonal evolution [19]. Within the ANR Fresqueau project we are studying tools for sequential pattern extraction taking into account spatial relations [56], [43].

From another point of view, miscanthus is a perennial crop used for biomass production. Its implantation is rather new, and there is few farms cultivating miscanthus in France. Understanding the farmers' choices for allocating miscanthus in their farmland is a main challenge. The CBR model is investigated for modeling these choices from farm surveys, including spatial reasoning aspects [20], [47] [41].

For completing the work on FCA and itemset search, there is still on-going work on frequent and rare itemset search, for being able to build lattices from very large data and completing the algorithm collection of the Coron platform. Work is still in progress on the design of an integrated and modular algorithm for searching for closed and generators itemsets, and equivalence classes of itemsets, thus enabling the construction of the associated lattice [121]. This research aspect is also linked to the research carried on within a the PICS CaDoE research project (see Section 8.1.1.3). In addition, there is also research work carried on different aspects involving the management of big data in the context of the BioIntelligence Project and the Quaero Project.

6.1.3. Skylines, sequential data, privacy and E-sports analytics

Pattern discovery is at the core of numerous data mining tasks. Although many methods focus on efficiency in pattern mining, they still suffer from the problem of choosing a threshold that influences the final extraction result. One goal is to make the results of pattern mining useful from a user-preference point of view. That is, take into account some domain knowledge to guide the pattern mining process. To this end, we integrate into the pattern discovery process the idea of skyline queries in order to mine *skyline patterns* in a threshold-free manner. This forms the basis for a novel approach to mining skyline patterns. The efficiency of our approach was illustrated over a use case from *chemoinformatics* and we showed that small sets of dominant patterns are produced under various measures that are interesting for chemical engineers and researchers.

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.

However, important fundamental problems still remained open: (i) given a sequence database, can we have an upper bound on the number of sequential patterns in the database? (ii) Is the efficiency of the sequence classifier only based on accuracy? (iii) Do the classifiers need the entire set of extracted patterns or a smaller set with the same expressiveness power?

In the field of the management of sequential date in medicine, analysis of health care trajectories led to the development of a new sequential pattern mining method [42]. The MMISP algorithm is able to efficiently extract sequential patterns composed of itemsets and multidimensional items. The multidimensional items can be described with additional taxonomic knowledge, allowing mining with appropriate levels of granularity. In parallel, a new measure has been created to compute the similarity between sequences of itemsets [78].

Orpailleur is one of the few project-teams working on privacy challenges which are becoming a core issue with different scientific problems in computer science. With technology infiltrating more and more every aspect of our lives, each human activity leaves a digital trace in some repository. Vast amounts of personal data are implicitly or explicitly created each day, and rarely one is aware of the extent of information that is kept, processed and analyzed without his knowledge or consent. These personal data give rise to significant concerns about user privacy, since important and sensitive details about private life are collected and exploited by third parties. The goal of privacy preservation technologies is to provide tools that allow greater control over the dissemination of user data. A promising trend in the field is Privacy Preserving Data Publishing (PPDP), which allows sharing of anonymized data. Anonymizing a dataset is not limited to the removal of direct identifiers that might exist in a dataset, e.g. the full name or the Social Security Number of a person. It also includes removing secondary information, e.g. like age, zip code that might lead indirectly to the true identity of an individual.

Existing research on this problem either perturbs the data, publishes them in disjoint groups disassociated from their sensitive labels, or generalizes their values by assuming the availability of a generalization hierarchy. In a recent work, we proposed a novel alternative [54]. Our publication method also puts data in a generalized form, but does not require that published records form disjoint groups and does not assume a hierarchy either. Instead, it employs generalized bitmaps and recasts data values in a nonreciprocal manner.

One of the most fascinating challenges of our time is understanding the complexity of the global interconnected society we inhabit. Today we have the opportunity to observe and measure how our society intimately works, by analyzing the big data. i.e, the digital breadcrumbs of human activities sensed as a by-product of the ICT

systems that we use. These data describe the daily human activities: for instance, automated payment systems record the tracks of our purchases, search engines record the logs of our queries for finding information on the web, social networking services record our connections to friends, colleagues and collaborators, wireless networks and mobile devices record the traces of our movements and our communications. These social data are at the heart of the idea of a knowledge society, where decisions can be taken on the basis of knowledge in these data.

Social network data analysis raises concerns about the privacy of related entities or individuals. We theoretically establish that any kind of structural identification attack can effectively be prevented using random edge perturbation and show that, surprisingly, important properties of the whole network, as well as of subgraphs thereof, can be accurately calculated and hence data analysis tasks performed on the perturbed data, given that the legitimate data recipient knows the perturbation probability as well [53].

"Electronic-sport" (E-Sport) is now established as a new entertainment genre. More and more players enjoy streaming their games, which attract even more viewers. In fact, in a recent social study, casual players were found to prefer watching professional gamers rather than playing the game themselves. Within this context, advertising provides a significant source of revenue to the professional players, the casters (displaying other people's games) and the game streaming platforms. In a recent work with Mehdi Kaytoue, we started focusing on the huge amount of data generated by electronic games. We crawled, during more than 100 days, the most popular among such specialized platforms: Twitch.tv.

Thanks to these gigabytes of data, we proposed a first characterization of a new Web community, and we showed, among other results, that the number of viewers of a streaming session evolves in a predictable way, that audience peaks of a game are explainable and that a Condorcet method can be used to sensibly rank the streamers by popularity [45]. This work should bring to light the study of E-Sport and its growing community for computer scientists and sociologists. They indeed deserve the attention of industrial partners (for the large amount of money involved) and researchers (for interesting problems in social network dynamics, personalized recommendation, sentiment analysis, etc.).

6.1.4. 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. This methodology is based on both FCA and RCA, and was previously successfully applied in contexts such as astronomy and biology. At present, an engineer is implementing a robust system being guided by the previous research results and preparing the way for some new research directions involving trees and graphs (see also the work on the ANR Hybride project).

6.2. KDDK in Life Sciences

Participants: Yasmine Assess, Emmanuel Bresso, Thomas Bourquard, Adrien Coulet, Marie-Dominique Devignes, Anisah Ghoorah, Renaud Grisoni, Jean-François Kneib, Florence Le Ber, Bernard Maigret, Jean-François Mari, Amedeo Napoli, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone.

The Life Sciences constitute a challenging domain in which to implement knowledge-guided approaches for knowledge discovery. Biological data are complex from many points of views: voluminous, high-dimensional, deeply inter-connected, etc. Analyzing such data and extracting hidden knowledge has become a crucial issue in important domains such as health, environment and agronomy. More and more bio-ontologies are available and can be used to enhance the knowledge discovery process [88], [117]. In the next few years, the experience of the Orpailleur team in KDDK applied to the Life Sciences will be further developed in two directions: the use of bio-ontologies to improve approaches for data integration and mining when applied to real-world data, and the study of the synergy between numeric and symbolic data-mining methods in life-science applications.

6.2.1. Relational data mining applied to complex biological object characterization and prediction

Inductive Logic Programming (ILP) is a learning method which allows expressive representation of the data and produces explicit first-order logic rules. However, any ILP system returns a single theory based on heuristic user-choices of various parameters and learning biases, thus ignoring potentially relevant rules. Accordingly, we propose an approach based on Formal Concept Analysis for effective interpretation of reached theories with the possibility of adding domain knowledge. Our approach was applied to the characterization of three-dimensional (3D) protein-binding sites, namely phosphorylation sites, which are the protein portions on which interactions with other proteins take place [33]. In this context, we defined a logical representation of 3D patches and formalized the problem as a concept learning problem using ILP. Another application of this KDDK methodology concerns the characterization and prediction of drug side-effect profiles (Journal manuscript in preparation). In this case, maximal frequent itemsets are extracted and allow us to propose relevant side-effect profiles of drugs which are further characterized by ILP.

6.2.2. Functional classification of genes using semantic similarity matrix and various clustering approaches

In the last report, we proposed a measure called IntelliGO which computes semantic similarity between genes for discovering biological functions shared by a set of genes (e.g., showing the same expression profile). This measure takes into account domain knowledge represented in Gene Ontology (GO) [83].

Functional classification aims at grouping genes according to their molecular function or the biological process they participate in. Evaluating the validity of such unsupervised gene classification remains a challenge given the variety of distance measures and classification algorithms that can be used. We evaluated functional classification of genes with the help of reference sets. Overlaps between clusters and reference sets are estimated by the F-score metric. We test the IntelliGO measure with hierarchical and fuzzy C-means clustering algorithms and we compare results with the state-of-the-art DAVID functional classification method (Database for Annotation Visualization and Integrated Discovery). Finally, study of best matching clusters to reference sets leads us to propose a method based on set-differences for discovering missing information.

The IntelliGO-based functional clustering method was tested on four benchmarking datasets consisting of biological pathways (KEGG database) and functional domains (Pfam database) [13]. The IntelliGO measure is usable on line (see http://bioinfo.loria.fr/Members/benabdsi/intelligo_project/).

We are currently investigating the clustering problem when objects are not represented as feature vectors in a vector space but as a pairwise similarity matrix. In biology such similarity measures are often computationally expensive or incompatible with *bona fide* distance definition. Embedding techniques of pairwise data into Euclidean space aim at facilitating subsequent clustering of the objects [115]. Spectral clustering methods are also relevant in this case [127]. We are conducting comparative and large-scale gene clustering evaluation using the Intelligo measure and reference sets.

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 tasks. In 2011 we collaborated with the National Center for Biomedical Ontologies (NCBO) to develop of large repository of annotations named the NCBO Resource Index [99]. The resulting repository contains annotations of 34 biomedical databases annotated with concepts of 280 ontologies of the BioPortal [2]. We proposed a comparison of the annotations of a database of biomedical publications (Medline) with two databases of scientific funding (Crisp and ResearchCrossroads) to profile disease research [18]. The annotation of these three databases with a unique ontology about diseases enable to consider their content conjointly and consequently to analyze and compare, for distinct disease (or family of diseases), trends in term of number of publications and funding amounts.

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

We started a new project that aims at exploring biomedical annotations with FCA techniques. One main challenge here is to develop a knowledge discovery approach that consider the knowledge represented in the ontologies employed for the annotations.

6.2.4. Connecting textual biomedical knowledge with the Semantic Web

A large amount of biomedical knowledge is in the form of text embedded in published articles, clinical files or biomedical public databases. It is consequently of high interest to extract and structure this knowledge to facilitate its consideration when processing biomedical data. We benefited from advances in Natural Language Processing (NLP) techniques to extract fine-grained relationships mentioned in biomedical text and subsequently published such relationships on line in the form of RDF triples [91], [90]. In a collaborative work with the Health Care and Life Science (HCLS) interest group of the W3C, we demonstrated how biomedical knowledge extracted from text, along with Semantic Web technologies has high potential for recommendation systems and knowledge discovery in biomedicine [118].

6.3. Structural Systems Biology

Participants: Thomas Bourquard, Marie-Dominique Devignes, Anisah Ghoorah, Van-Thai Hoang, Bernard Maigret, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone.

knowledge discovery in life sciences, bioinformatics, biology, chemistry, gene

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) [105]. The docking server has performed some 12,000 docking runs during 2012. A book chapter describing the Hex docking algorithm has been published [75]. Our docking work has facilitated further developments on modeling the assembly of multi-component molecular structures using a particle swarm optimization technique [25], and on modeling protein flexibility during docking [24].

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. KBDOCK combines residue contact information from the 3DID database [119] with the Pfam protein domain family classification [92] together with coordinate data from the Protein Data Bank [87] 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 [96] for template-based docking using 73 complexes from the Protein Docking Benchmark [98]. KBDOCK is available at http://kbdock.loria.fr. Anisah Ghoorah successfully defended her thesis in November 2012 [10].

6.3.3. Kpax: A new algorithm for protein structure alignment

We have developed a new protein structure alignment approach called Kpax [6]. 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 [126]. The Kpax program is available at http://kpax.loria.fr/.

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. 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. Both programs will soon be made publicly available, and two manuscripts describing our approach are in preparation.

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 is 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 [111]. The main disease target in this project is the acquired immune deficiency syndrome (AIDS), caused by the human immuno-deficiency virus (HIV) [109]. However, the approach will be quite generic and will be broadly applicable to many other diseases. Good progress has been made on calculating and clustering spherical harmonic "consensus shapes" which represent rather well the essential features of groups of active molecules [110]. The approach has since been extended to provide a rapid way to cluster drug families according to the Gaussian distributions of their surface shapes, and to predict possible cross-interactions of drug families [21]. We have also published a review on the state of the art in 3D virtual drug screening [15].

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.

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

The Taaable project (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, information retrieval, knowledge acquisition and extraction, knowledge representation, minimal change theory, ontology engineering, semantic wikis, text-mining, etc. Case-based reasoning 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.

The most important original features of this version are:

Modules for computing adaptation knowledge. Using adaptation knowledge, and especially adaptation rules, is a way to better adapt cooking recipes to user constraints. A previous work for extracting adaptation rules has been performed in 2011 [93]. In this work, variation of ingredients between couple of recipes are mined using closed itemsets extraction. The adaptation rules come from the interpretation of closed itemsets whose items correspond to the ingredients that have to be removed, kept, or added. This approach has been integrated as a wiki extension, providing a collaborative environment in which humans and machines may now collaborate to better acquire adaptation rules [38]. Humans (expert in cooking) may trigger automatic processes (knowledge discovery processes) and and may validate, using a specific user interface, proposition of adaptation rules as adaptation knowledge, which is then added to the knowledge base. In the same way, this environment integrates also the results of a new work on knowledge extraction where specific cooking adaptation rules (i.e. that can be applied to a single recipe) are generalized using close itemsets into generic adaptation rules, to make them usable on other recipes [60].

A module for acquiring a process semantic representation. While a process for acquiring cases from recipe preparation texts exists, the results are not perfect. In order for valid case representations to be available in the semantic wiki, a semi-automatic case acquisition tool was created [59]. This tools presents the user with a graphical interface through which it is able to interact with the case acquisition process. In order to limit the effort required, each correction entered by the user is propagated by the tool to the rest of the case representation.

Some other theoretical studies have been carried out that should be applied to some future versions of Taaable:

- The combination of workflows and interval algebras to represent procedural knowledge [55].
- The revision-based adaptation of cases represented in a qualitative algebra [41].
- The study of taxonomy merging [39]: several versions of the taxonomies used in Taaable (such as the food hierarchy) can be incoherent one with the others and a merging process is defined in order to obtain a consistent merged taxonomy.
- A continuous knowledge extraction process to ensure the non regression of the reasoning system according to the ontology evolution [50].

7. Bilateral Contracts and Grants with Industry

7.1. The BioIntelligence Project

Participants: Mehwish Alam, Yasmine Assess, Aleksey Buzmakov, Adrien Coulet, Marie-Dominique Devignes, Amedeo Napoli [contact person], 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 running in the Orpailleur team. A first thesis is related to the study of possible combination of mining methods on biological data. The mining methods which are considered here are based on FCA and RCA, itemset and association rule extraction, and inductive logic programming. These methods have their own strengths and provide different special capabilities for extending domain ontologies. A particular attention will be paid to the integration of heterogeneous biological data and the management of a large volume of biological data while being guided by domain knowledge lying in ontologies (linking data and knowledge units). 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 Graphs. The idea is to be able to extend the formalism of pattern structures to graphs and to apply the resulting framework on molecular structures. In this way, it will be possible to classify molecular structures and reactions by their content. This will help practitioners in information retrieval tasks involving molecular structures or the search for particular reactions. In addition, an experiment was also carried out in the combination of supervised (distance-based clustering) and unsupervised learning (FCA) methods for the prediction of the configuration of inhibitors of the c-Met protein (which is very active in cancer).

In addition, a forthcoming thesis will be in concern with ontology re-engineering in the domain of biology. The objective is consider the content of the BioPortal ontologies (http://bioportal.bioontology.org/) and to design formal contexts and associated concept lattices which will become supports for ontological schemes. Moreover, this ontological schema will be completed thanks to external resources such as Wikipedia and domain knowledge as well. The global idea is to get definitions and thus classification capabilities for atomic or primitive concepts.

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 participates in the task called “Formal Representation of Knowledge for Guiding Recommendation”, whose objectives are to define methods and algorithms for building a “discovery engine” guided by domain knowledge and able to recommend a user some content to visualize. Such a discovery engine has to extend capabilities of usual recommender systems with a number of capabilities, e.g. to select among a huge amount of items (e.g. movie, video, music) those which are of interest for a user according to a given profile. In addition, the discovery engine should take into account contextual information that can be of interest such as news, space location, moment of the day, actual weather and weather forecast, etc. This contextual information changes within time and extracted information has to be continuously updated. Finally, the system has be able to justify or explain the recommendations.

A thesis takes place in the context of the Quaero project. At the moment, document annotation is especially studied for enhancing recommendation but also information retrieval. Information retrieval guided by domain knowledge can be used for selecting resources of interest for these two tasks. Then knowledge discovery based on Formal Concept Analysis can be used for extracting patterns of interest w.r.t. the context and for enriching the domain and contextual knowledge base.

Finally, the discovery process has to be able to act as a classifier and as an inference engine at the same time for reasoning and classifying elements for recommendation and retrieval.

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.

Combining Numerical and Symbolical Methods for the Classification of Multi-valued and Interval Data (CM2ID)

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, in order to push the complexity barrier farther in the mining of complex data. Biological data, namely gene expression data, are used for test and evaluation of the combination of algorithms.

The project involves three teams, one Brazilian team and two French Inria teams, including specialists of clustering and classification methods. Thus the complementarity of the teams is ensured and, in addition, close contacts exist with experts of the domain of data for carrying on a complete evaluation of the results obtained by the combined algorithms expected to be designed during the project.

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.

One strategy to ease the execution of such data mining tasks is to use existing knowledge to restrict the search space and to assess the quality of the patterns found. This existing knowledge may be formalized in ontologies but also in other ways whose study is a research issue in this project. Once we are able to build knowledge models, we need to determine how to use such knowledge models, which is a second major research issue in this project. In particular, we want to design and evaluate mechanisms that allow the exploitation of existing knowledge for sake of improving data mining algorithms.

Finally, the computational complexity of the algorithms remains a major issue and we intend to address it through parallel algorithms. Data mining algorithms, in general, represent a challenge for sake of parallelization because they are irregular and intensive in terms of both computing and communication. Accordingly, in a first joint work, we developed a new parallel algorithm to build skycubes based on the Anthill framework developed at UFMG. The paper was presented in a local Brazilian Conference and an extended journal version will appear in a 2012 special issue of the International Journal of Parallel Programming.

8.1.1.3. International collaborations in Mining complex data

Participants: Mehwish Alam, Aleksey Buzmakov, Victor Codocedo, Adrien Coulet, Elias Egho, Ioanna Lykourentzou, Amedeo Napoli [contact person], Chedy Raïssi.

8.1.1.3.1. PICS CNRS CADOE

A first collaboration involves "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). At the moment, some elements of this methodology are existing and were used in text mining [86], [85], but this methodology should be completed and improved, especially regarding the applicability on complex data and the interoperability with knowledge representation modules.

8.1.1.3.2. Collaboration with HSE Moscow

A second collaboration involves Sergei Kusnetsov at Higher School of Economics in Moscow (HSE). Amedeo Napoli visited HSE laboratory in November 2012 (with the support of HSE) and Sergei Kuznetsov visited Inria NGE in August and in December 2012. These visits were the occasion of preparing a publications (submitted for the next year). This shows that the collaboration is on-going and that there is still a substantial research work to be done.

8.1.1.3.3. AGAUR Project: collaboration with UPC Barcelona

This project mainly involves Amedeo Napoli and Jaume Baixeries who is an Associate Professor at UPC Barcelona (Universitat Politècnica de Catalunya). Amedeo Napoli had a stay of roughly two months in December 2011 and May-June 2012. Both researchers have worked, jointly with Mehdi Kaytoue, on the characterization of functional dependencies in many-valued data with FCA and pattern structures. In this work, functional dependencies are directly taken into account and this shows a different but important capability of pattern structures to deal with complex data [30].

8.1.1.3.4. PHC Zenon (Cyprus)

A third collaboration –a PHC Zenon project– exists with Florent Domenach, associated professor at the University of Nicosia in Cyprus. This project is 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. These processes can be based on Galois connections but also on the so-called "overhangings", i.e. a kind of generalization of closure systems. Moreover, another interest is put on consensus theory where the objective is to find the better classification of a set of abjects according to a quality measure (this could be applied to ontologies). This year, there were two visits, one from Cyprus to France in October 2012 and the other from France to Cyprus in December 2012. Publications are currently submitted.

8.2. European Initiatives

8.2.1. FP7 Project DOVSA

DOVSA stands for "Development of Virtual Screening Algorithms: Exploring Multiple Ligand Binding Modes Using Spherical Harmonic Consensus Clustering". It is a European project (Type PEOPLE) funded as a "Marie Curie Intra-European Fellowships for Career Development (IEF)" from July 2010 until July 2012. The coordinator of the project is Inria NGE.

This project is aimed at advancing the state of the art in virtual drug screening by developing novel spherical harmonic-based consensus clustering algorithms. The main disease that will be targeted in this project is the acquired immune deficiency syndrome (AIDS), caused by the human immuno-deficiency virus (HIV). However, the approach will be quite generic and will be broadly applicable to many other diseases. The approach will be tested and validated using 40 well-known drug targets from the DUD dataset. It will then be used to screen the French Chimiothèque Nationale library of some 36000 compounds for novel ligands which will bind the CCR5 co-receptor and hence block HIV infection. A small list of candidate entry-blocking compounds will be sent to Barcelona for experimental testing. By extending the SH-based consensus clustering

technique, this project will provide a generic tool to help deal with cases where multiple ligands may be associated with multiple pocket sub-sites or which may bind multiple targets, and it will help to find new HIV entry-blocking compounds.

8.3. National Initiatives

8.3.1. ANR

8.3.1.1. ANR Hybride

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

The Hybride research project 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, while 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. This combination of NLP and KDD methods for common objectives, can be viewed as a continuous process, based on a sequence of complex operations from NLP and KDD that reinforces itself through a feedback loop. 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 fundamental aspects of the Hybride project can be understood through the main steps of the knowledge discovery loop with a NLP/KDD perspective : (i) data preparation, (ii) data mining, (iii) interpretation and validation of the results, (iv) knowledge construction. At each step, new methods have to be designed for achieving this interrelated NLP/KDD loop. One of the outcomes of the project should be a system integrating the operations involved within the whole NLP/KDD loop, in the context of Orphanet for text analysis and production of new documentation on rare diseases. The implementation of such a system combines various interrelated aspects, namely natural language processing, knowledge discovery, data mining, and knowledge engineering. This original combination still remains a challenge in computer science.

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 Inria NGE (FCA, knowledge representation, pattern mining, text mining).

8.3.1.2. ANR Kolflow

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

Kolflow (http://kolflow.univ-nantes.fr/) is a 3-years 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, Score (LORIA), and Wimmics (Inria Sophia Antipolis).

8.3.1.3. ANR PEPSI: Polynomial Expansions of Protein Structures and Interactions

Participants: Dave 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 programme 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).

8.3.1.4. ANR 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 knowledge for other 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). 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 id to provide in decision support for improving healthcare in detecting for example typical or exceptional trajectories for planning with precision healthcare for a given population. In order to discover groups of patients showing similar health condition, treatments or journeys through the healthcare system, PCT are modeled as multilevel and multidimensional sequences of itemsets, using external knowledge on hospitals, medical procedures and diagnoses. Accordingly, a new algorithm [42] has been developed to mine sequential patterns.

8.3.2. Other National Inititives and Collaborations

8.3.2.1. PEPS Cryo-CA

Participant: Dave 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 Dave 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. In November 2012, a workshop funded by this project attracted some 60 participants (http://ccsb2012.loria.fr).

8.3.2.2. 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 collaborate on the NRPS toolbox [57]. 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.

8.4. Regional Initiatives

8.4.1. BioProLor

The Orpailleur team is member of the BioProLor consortium composed of 5 enterprises and 7 academic research teams. This consortium is funded for 2 years (2010-2012) by the AME ("Agence pour la Mobilisation Economique"). The objective of BioProLor is the design of a production filière for compounds with high added-value which originate from plants in Lorraine. The Orpailleur team and the associated start-up "Harmonic Pharma" are in charge of the computational aspects of this research work.

In addition, a CIFRE contract (2009-2012) was set up with Harmonic Pharma for funding the thesis of Emmanuel Bresso on the following subject: “Organisation et exploitation des connaissances sur les réseaux d’interactions biomoléculaires pour l’identification de gènes candidats et la caractérisation de profils d’effets secondaires de principes actifs”.

8.4.2. Contrat Plan État Région” (CPER)

The links between the Regional Administration and LORIA are materialized through an administrative contract called “Contrat Plan État Région” (CPER) 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).

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

  An Inria experimental research platform is currently developed in the framework of MBI (http://bioinfo.loria.fr/Plateforme%20MBI). This platform is aimed at sharing data and computing resources. Its specific features are relative to biomolecules modeling, classification, and to data integration for data mining. In parallel with the bioinformatics platforms in Strasbourg, Reims, Lille, and Nancy-INIST, it constitutes the North-East node of RENABI (“Réseau National des Plateformes Bioinformatiques”).

- “Traitement Automatique des Langues et des Connaissances” (TALC).

  TALC stands for “Automatic Processing of Languages and Knowledge”. The general objective is to study the relations existing between knowledge discovery, knowledge representation, reasoning, and natural language processing. In this framework, the Orpailleur team plays an important role as the research themes are closely related to those of the team. Actually, research projects are currently under development on knowledge management and decision support in the large 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 at least twice a month and is used either for general presentations of people in the team or for inviting external researchers for general interest. The BINGO seminar is held also at least 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, as members of conference program committees (ECAI, IJCAI, PKDD, ICFCA ...), as members of editorial boards, and finally in the organization of journal special issues.
- This year, Dave Ritchie co-organized a workshop on Computational Challenges in Structural Biology (CCSB-2012; http://ccsb2012.loria.fr).

- Emmanuel Nauer co-organized a new workshop, called "Cooking with Computers" at ECAI 2012 (Montpellier). This workshop aims at bringing together researchers from every possible fields of artificial intelligence applying their research on food and cooking. The next workshop will take place in 2013 at the ICJAI Conference (August 2013, Beijing, China).
- Amedeo Napoli was one of the co-organizers of the ECAI workshop FCA4AI (What Artificial Intelligence can do for FCA), together with Sergei O. Kuznetsov and Sebastian Rudolph (see the website of the workshop http://www.fca4ai.hse.ru and the CEUR proceedings http://ceur-ws.org/Vol-939/). Based on the large success of this first workshop, a new edition will take place at the ICJAI Conference in August 2013, Beijing, China.
- Chedy Raïssi was a co-organizer of the “PinSoDa” workshop, joint with the 11th IEEE International Conference on Data Mining (ICDM 2012), Brussels, Belgium. The purpose of this workshop was to encourage principled research that will lead to the advancement of the science of privacy and data protection on social data.

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 (in Nancy for most of them). 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] Y. ASSES, V. VENKATRAMAN, V. LEROUX, D. RITCHIE, B. MAIGRET. *Exploring c-Met kinase flexibility by sampling and clustering its conformational space*, in "Proteins", January 2012, vol. 80, n^{o} 4, pp. 1227-1238 [*DOI :* 10.1002/PROT.24021], http://hal.inria.fr/hal-00756791

[2] E. BRESSO, R. GRISONI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Formal Concept Analysis for the Interpretation of Relational Learning applied on 3D Protein-Binding Sites*, in "4th international conference on Knowledge Discovery and Information Retrieval - KDIR 2012", Barcelona, Spain, A. FRED (editor), 2012, 12 pages p. , http://hal.inria.fr/hal-00734349

[3] 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

[4] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER, L. MARTIN. *Adapting Spatial and Temporal Cases*, in "International Conference for Case-Based Reasoning", Lyon, France, I. WATSON, B. D. AGUDO (editors), Lecture Notes in Artificial Intelligence, Springer, September 2012, vol. 7466, pp. 77-91 [*DOI :* 10.1007/978-3-642-32986-9_8], http://hal.inria.fr/hal-00735231

[5] 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° e1, pp. e177-e186 [*DOI :* 10.1136/AMIAJNL-2011-000631], http://hal.inria.fr/hal-00752101

[6] 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° 24, pp. 3274-3281 [*DOI :* 10.1093/BIOINFORMATICS/BTS618], http://hal.inria.fr/hal-00756813

[7] 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° 3, pp. 433-446 [*DOI :* 10.1007/s10980-011-9691-2], http://hal.inria.fr/hal-00656407

[8] H. SKAF-MOLLI, E. DESMONTILS, E. NAUER, G. CANALS, A. CORDIER, M. LEFEVRE, P. MOLLI, Y. TOUSSAINT. *Knowledge Continuous Integration Process (K-CIP)*, in "WWW 2012 - SWCS'12 Workshop - 21st World Wide Web Conference - Semantic Web Collaborative Spaces workshop", Lyon, France, April 2012, pp. 1075-1082, http://hal.inria.fr/hal-00765596

[9] M. XUE, P. KARRAS, C. RAÏSSI, J. VAIDYA, K.-L. TAN. *Anonymizing set-valued data by nonreciprocal recoding*, in "The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12,", Beijing, China, Q. YANG, D. AGARWAL, J. PEI (editors), August 2012, http://hal.inria.fr/hal-00768428

Publications of the year

Doctoral Dissertations and Habilitation Theses

[10] 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

Articles in International Peer-Reviewed Journals

[11] Y. ASSES, V. VENKATRAMAN, V. LEROUX, D. RITCHIE, B. MAIGRET. *Exploring c-Met kinase flexibility by sampling and clustering its conformational space*, in "Proteins", January 2012, vol. 80, n° 4, pp. 1227-1238 [*DOI :* 10.1002/PROT.24021], http://hal.inria.fr/hal-00756791

[12] A. COULET, K. B. COHEN, R. B. ALTMAN. *Guest Editorial: The state of the art in text mining and natural language processing for pharmacogenomics*, in "Journal of Biomedical Informatics", October 2012, vol. 45, n° 5, pp. 825-826, http://hal.inria.fr/hal-00752106

[13] 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° 3/4, pp. 245-260, http://hal.inria.fr/hal-00734329

[14] A. FURLAN, F. COLOMBO, A. KOVER, N. ISSALY, C. TINTORI, L. ANGELI, V. LEROUX, S. LETARD, M. AMAT, Y. ASSES, B. MAIGRET, P. DUBREUIL, M. BOTTA, R. DONO, J. BOSCH, O. PICCOLO, D. PASSARELLA, F. MAINA. *Identification of new aminoacid amides containing the imidazo[2,1-b]benzothiazol-2-ylphenyl moiety as inhibitors of tumorigenesis by oncogenic Met signaling*, in "European Journal of

Medicinal Chemistry", 2012, vol. 47, pp. 239 - 254 [*DOI :* 10.1016/J.EJMECH.2011.10.051], http://hal.inria.fr/hal-00763849

[15] L. GHEMTIO, V. PÉREZ-NUENO, V. LEROUX, Y. ASSES, M. SOUCHET, L. MAVRIDIS, B. MAIGRET, D. RITCHIE. *Recent Trends and Applications in 3D Virtual Screening*, in "Combinatorial Chemistry and High Throughput Screening", August 2012, vol. 15, n^{o} 9, pp. 749-769, http://hal.inria.fr/hal-00756800

[16] C. GOETZ, A. ZANG, N. JAY. *Apports d'une méthode de fouille de données pour la détection des cancers du sein incidents dans les données du programme de médicalisation des systèmes d'information*, in "Informatique et Sante", 2012, vol. 1, pp. 189-199 [*DOI :* 10.1007/978-2-8178-0285-5_17], http://hal.inria.fr/hal-00764959

[17] 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

[18] 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, pp. e177-e186 [*DOI :* 10.1136/AMIAJNL-2011-000631], http://hal.inria.fr/hal-00752101

[19] S. MARTIN, A. BERTAUX, F. LE BER, E. MAILLARD, G. IMFELD. *Seasonal Changes of Macroinvertebrate Communities in a Stormwater Wetland Collecting Pesticide Runoff From a Vineyard Catchment (Alsace, France)*, in "Archives of Environmental Contamination and Toxicology", 2012, vol. 62, n^{o} 1, pp. 29-41 [*DOI :* 10.1007/S00244-011-9687-6], http://hal.inria.fr/hal-00607741

[20] L. MARTIN, J. WOHLFAHRT, F. LE BER, M. BENOÎT. *L'insertion territoriale des cultures biomasses pérennes. Etude de cas sur le miscanthus en Côte d'Or (bourgogne, France)*, in "L'Espace Géographique", 2012, n^{o} 2, pp. 138-153, http://hal.inria.fr/hal-00733672

[21] V. PÉREZ-NUENO, V. VENKATRAMAN, L. MAVRIDIS, D. RITCHIE. *Detecting Drug Promiscuity Using Gaussian Ensemble Screening*, in "Journal of Chemical Information and Modeling", July 2012, vol. 52, n^{o} 8, pp. 1948-1961 [*DOI :* 10.1021/CI3000979], http://hal.inria.fr/hal-00756804

[22] 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

[23] 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

[24] 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

[25] 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

Invited Conferences

[26] C. RAÏSSI. *Multidimensional skylines*, in "Actes des 8èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, EDA 2012", Bordeaux, France, S. MAABOUT (editor), October 2012, http://hal.inria.fr/hal-00768448

International Conferences with Proceedings

[27] M. ALAM, A. COULET, A. NAPOLI, M. SMAÏL-TABBONE. *Formal Concept Analysis Applied to Transcriptomic Data*, in "What can FCA do for Artificial Intelligence (FCA4AI, ECAI 2012 Workshop)", Montpellier, France, S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors), August 2012, http://hal.inria.fr/hal-00760993

[28] M. ALAM, A. COULET, A. NAPOLI, M. SMAÏL-TABBONE. *Formal Concept Analysis Applied to Transcriptomic Data*, in "The Ninth International Conference on Concept Lattices and Their Applications - CLA 2012", Fuengirola (Málaga), Spain, L. SZATHMARY, U. PRISS (editors), October 2012, http://hal.inria.fr/hal-00761003

[29] Y. ASSES, A. BUZMAKOV, T. BOURQUARD, S. O. KUZNETSOV, A. NAPOLI. *A Hybrid Classification Approach based on FCA and Emerging Patterns - An application for the classification of biological inhibitors*, in "CLA'12: The Ninth International Conference on Concept Lattices and Their Applications - 2012", Fuengirola, Spain, L. SZATHMARY, U. PRISS (editors), October 2012, http://hal.inria.fr/hal-00761586

[30] J. BAIXERIES, M. KAYTOUE, A. NAPOLI. *Computing Functional Dependencies with Pattern Structures*, in "The 9th International Conference on Concept Lattices and Their Applications - CLA 2012", Malaga, Spain, L. SZATHMARY, U. PRISS (editors), October 2012, http://hal.inria.fr/hal-00763748

[31] S. BEN ABBÈS, A. SCHEUERMANN, T. MEILENDER, M. D'AQUIN. *Characterizing Modular Ontologies*, in "7th International Conference on Formal Ontologies in Information Systems - FOIS 2012", Graz, Austria, July 2012, pp. 13-25, http://hal.inria.fr/hal-00710035

[32] A. BERRY, M. HUCHARD, A. NAPOLI, A. SIGAYRET. *Hermes: an efficient algorithm for building Galois sub-hierarchies*, in "CLA'2012: 9th International Conference on Concept Lattices and Applications", Fuengirola (Málaga), Spain, L. SZATHMARY, U. PRISS (editors), Universidad de Malaga, October 2012, pp. 21-32, http://hal.inria.fr/lirmm-00743882

[33] E. BRESSO, R. GRISONI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Formal Concept Analysis for the Interpretation of Relational Learning applied on 3D Protein-Binding Sites*, in "4th international conference on Knowledge Discovery and Information Retrieval - KDIR 2012", Barcelona, Spain, A. FRED (editor), 2012, 12 pages p. , http://hal.inria.fr/hal-00734349

[34] A. BUZMAKOV, S. O. KUZNETSOV, A. NAPOLI. *A New Approach to Classification by Means of Jumping Emerging Patterns*, in "FCA4AI: International Workshop "What can FCA do for Artificial Intelligence?" - ECAI 2012", Montpellier, France, S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors), August 2012, http://hal.inria.fr/hal-00761602

[35] V. CODOCEDO, I. LYKOURENTZOU, A. NAPOLI. *A contribution to semantic indexing and retrieval based on FCA - An application to song datasets*, in "Proceedings of the conference on concept lattices and their applications (CLA)", Malaga, Spain, L. SZATHMARY, U. PRISS (editors), October 2012, http://hal.inria.fr/hal-00760764

[36] V. CODOCEDO, I. LYKOURENTZOU, A. NAPOLI. *Semantic querying of data guided by Formal Concept Analysis*, in "Proceedings of the ECAI Workshop on Formal Concept Analysis for Artificial Intelligence (FCA4AI)", Montpellier, France, S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors), December 2012, http://hal.inria.fr/hal-00760757

[37] J. COJAN, J. LIEBER. *Belief revision-based case-based reasoning*, in "ECAI-2012 Workshop SAMAI", Montpellier, France, G. RICHARD (editor), 2012, pp. 33–39, http://hal.inria.fr/hal-00763220

[38] A. CORDIER, E. GAILLARD, E. NAUER. *Man-Machine Collaboration to Acquire Cooking Adaptation Knowledge for the TAAABLE Case-Based Reasoning System*, in "SWCS Semantic Web Collaborative Spaces", Lyon, France, ACM, April 2012, pp. 1113-1120, http://hal.inria.fr/hal-00696013

[39] A. CORDIER, J. LIEBER, J. STEVENOT. *Towards an operator for merging taxonomies*, in "ECAI-2012 Workshop BNC: Belief change, Non-monotonic reasoning and Conflict resolution", Montpellier, France, S. KONIECZNY, T. MEYER (editors), August 2012, http://hal.inria.fr/hal-00763228

[40] S. DA SILVA, C. LAVIGNE, F. LE BER. *Analyse des structures des haies et des linéaires pérennes dans deux paysages agricoles contrastés*, in "SAGEO - International Conference on Spatial Analysis and GEOmatics", Liège, Belgium, November 2012, 16 p. , http://hal.inria.fr/hal-00742681

[41] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER, L. MARTIN. *Adapting Spatial and Temporal Cases*, in "International Conference for Case-Based Reasoning", Lyon, France, I. WATSON, B. D. AGUDO (editors), Lecture Notes in Artificial Intelligence, Springer, September 2012, vol. 7466, pp. 77-91 [*DOI :* 10.1007/978-3-642-32986-9_8], http://hal.inria.fr/hal-00735231

[42] E. EGHO, D. IENCO, N. JAY, A. NAPOLI, P. PONCELET, C. QUANTIN, C. RAÏSSI, M. TEISSEIRE. *Healtcare Trajectory Mining by Combining Multi-dimensional Component and Itemsets*, in "NFMCP'2012: New Frontiers in Mining Complex Patterns, Workshop in conjunction with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012)", United Kingdom, LNCS, Springer, 2012, 12 p. , http://hal.inria.fr/lirmm-00732661

[43] M. FABRÈGUE, A. BRAUD, S. BRINGAY, F. LE BER, M. TEISSEIRE. *Including spatial relations and scales within sequential pattern extraction*, in "DS'2012: 15th International Conference on Discovery Science", Lyon, France, LNCS / LNAI, October 2012, http://hal.inria.fr/lirmm-00735617

[44] T. V. HOANG, S. TABBONE. *Fast Computation of Orthogonal Polar Harmonic Transforms*, in "The 21st International Conference on Pattern Recognition - ICPR 2012", Tsukuba Science City, Japan, November 2012, http://hal.inria.fr/hal-00734307

[45] M. KAYTOUE, A. SILVA, L. CERF, W. MEIRA, C. RAÏSSI. *Watch me playing, I am a professional: a first study on video game live streaming*, in "MSND@WWW - International Workshop on Mining Social Network Dynamics - 2012 (in conjunction with WWW - World Wild Web - 2012)", Lyon, France, H. HACID, S. GUO, J. VELCIN (editors), ACM, April 2012, pp. 1181-1188 [*DOI :* 10.1145/2187980.2188259], http://hal.inria.fr/hal-00697150

[46] F. LE BER, C. LAVIGNE, S. DA SILVA. *Structure analysis of hedgerows and other perennial landscape lines in two French agricultural landscapes*, in "AGILE 2012", Avignon, France, 2012, 6 p. , http://hal.inria.fr/hal-00685777

[47] L. MARTIN, F. LE BER, J. WOHLFAHRT, G. BOCQUÉHO, M. BENOÎT. *Modelling farmers' choice of miscanthus allocation in farmland: a case-based reasoning model*, in "2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet, Sixth Biennial Meeting", Leipzig, Germany, R. SEPPELT, A. VOINOV, S. LANGE, D. BANKAMP (editors), International Environmental Modelling and Software Society (iEMSs), July 2012, 8 p. , http://hal.inria.fr/hal-00724085

[48] T. MEILENDER, J. LIEBER, G. HERENGT, F. PALOMARES, N. JAY. *A Semantic Wiki for Editing and Sharing Decision Guidelines in Oncology*, in "The 24th European Medical Informatics Conference - MIE 2012", Pise, Italy, J. MANTAS (editor), M.Cristina Mazzoleni, August 2012, http://hal.inria.fr/hal-00736711

[49] T. MEILENDER, J. LIEBER, F. PALOMARES, N. JAY. *From Web 1.0 to Social Semantic Web: Lessons Learnt from a Migration to a Medical Semantic Wiki*, in "9th Extended Semantic Web Conference - ESWC 2012", Heraklion, Greece, May 2012, http://hal.inria.fr/hal-00736706

[50] H. SKAF-MOLLI, E. DESMONTILS, E. NAUER, G. CANALS, A. CORDIER, M. LEFEVRE, P. MOLLI, Y. TOUSSAINT. *Knowledge Continuous Integration Process (K-CIP)*, in "WWW 2012 - SWCS'12 Workshop - 21st World Wide Web Conference - Semantic Web Collaborative Spaces workshop", Lyon, France, April 2012, pp. 1075-1082, http://hal.inria.fr/hal-00765596

[51] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *Efficient Vertical Mining of Minimal Rare Itemsets*, in "The Nineth International Conference on Concept Lattices and their Applications - CLA 2012", Malaga, Spain, U. PRISS, L. SZATHMARY (editors), University of Malaga (Spain), 2012, pp. 269–280, http://hal.inria.fr/hal-00769031

[52] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *Finding minimal rare itemsets in a depth-first manner*, in "ECAI Workshop on Formal Concept Analysis for Artificial Intelligence (FCA4AI)", Montpellier, France, S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors), 2012, pp. 71–78, http://www.fca4ai.hse.ru, http://hal.inria.fr/hal-00769028

[53] M. XUE, P. KARRAS, C. RAÏSSI, P. KALNIS, H. K. PUNG. *Delineating social network data anonymization via random edge perturbation*, in "21st ACM International Conference on Information and Knowledge Management, CIKM'12", Maui, United States, X. WEN CHEN, G. LEBANON, H. WANG, M. J. ZAKI (editors), October 2012, http://hal.inria.fr/hal-00768441

[54] M. XUE, P. KARRAS, C. RAÏSSI, J. VAIDYA, K.-L. TAN. *Anonymizing set-valued data by nonreciprocal recoding*, in "The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12,", Beijing, China, Q. YANG, D. AGARWAL, J. PEI (editors), August 2012, http://hal.inria.fr/hal-00768428

National Conferences with Proceedings

[55] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER. *Extension du formalisme des flux opérationnels par une algèbre temporelle*, in "Sixièmes Journées de l'Intelligence Artificielle Fondamentale (JIAF)", Toulouse, France, Sébastien Konieczny, May 2012, pp. 133-142, http://hal.inria.fr/hal-00712978

[56] M. FABRÈGUE, A. BRAUD, S. BRINGAY, F. LE BER, M. TEISSEIRE. *Extraction de motifs spatio-temporels à différentes échelles avec gestion de relations spatiales qualitatives.*, in "Inforsid 2012", France, May 2012, pp. 123-138, http://hal.inria.fr/lirmm-00735616

[57] 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

Conferences without Proceedings

[58] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER, L. MARTIN. *Adaptation de cas spatiaux et temporels*, in "20ème atelier Français de Raisonnement à Partir de Cas", Paris, France, Laura Martin and Zied Yakoubi, June 2012, http://hal.inria.fr/hal-00712982

[59] V. DUFOUR-LUSSIER, F. LE BER, J. LIEBER, T. MEILENDER, E. NAUER. *Semi-automatic annotation process for procedural texts: An application on cooking recipes*, in "Cooking with Computers workshop - ECAI 2012", Montpellier, France, A. CORDIER, E. NAUER (editors), August 2012, http://hal.inria.fr/hal-00735262

[60] E. GAILLARD, E. NAUER, M. LEFEVRE, A. CORDIER. *Extracting Generic Cooking Adaptation Knowledge for the TAAABLE Case-Based Reasoning System*, in "Cooking with Computers workshop @ ECAI 2012", Montpellier, France, August 2012, http://hal.inria.fr/hal-00720481

[61] T. MEILENDER, J. LIEBER, F. PALOMARES, N. JAY. *Semantic decision trees editing for decision support with KcatoS - Application in oncology*, in "SIMI 2012: Semantic Interoperability in Medical Informatics", Heraklion, Greece, May 2012, http://hal.inria.fr/hal-00736710

[62] J.-H. RAMAROSON, D. HERVÉ, B. RAMAMONJISOA, F. LE BER. *Organisation spatiale, dessinée par les paysans, du corridor forestier de Fianarantsoa*, in "4ème édition Forum de la recherche: " Innovations scientifiques et technologiques - Valorisation de la recherche "", Antananarivo, Madagascar, July 2012, http://hal.inria.fr/hal-00764152

[63] J. WIEDERKEHR, B. FONTAN, C. GRAC, F. LABAT, F. LE BER, M. TRÉMOLIÈRES. *Stream multi-index assessment and associated uncertainties : application to macroinvertebrate and macrophytes*, in "Journées Internationales de Limnologie et d'Océanographie - JILO 2012", Clermont-Ferrand, France, October 2012, http://hal.inria.fr/hal-00764287

Scientific Books (or Scientific Book chapters)

[64] B. BUCHER, F. LE BER. , *Développements logiciels en géomatique*, Hermes Lavoisier, 2012, 297 p. , http://hal.inria.fr/hal-00724084

[65] B. BUCHER, F. LE BER. , *Innovative Software Development in GIS*, ISTE - WILEY, 2012, 331 p. , http://hal.inria.fr/hal-00724080

[66] B. BUCHER, J. GAFFURI, F. LE BER, T. LIBOUREL. *Challenges and Proposals for Software Development Pooling in Geomatics*, in "Innovative Software Development in GIS", B. BUCHER, F. LE BER (editors), GIS Series, ISTE - WILEY, 2012, pp. 293-316, ISBN : 978-1848213647, http://hal.inria.fr/hal-00724082

[67] B. BUCHER, J. GAFFURI, F. LE BER, T. LIBOUREL. *Défis et propositions pour la mutualisation de développements logiciels en géomatique*, in "Développements logiciels en géomatique – innovations et mutualisation", B. BUCHER, F. LE BER (editors), IGAT, Hermes Lavoisier, 2012, pp. 271-290, http://hal.inria.fr/hal-00718768

[68] B. BUCHER, F. LE BER. *Introduction*, in "Développements logiciels en géomatique – innovations et mutualisation", B. BUCHER, F. LE BER (editors), IGAT, Hermes Lavoisier, 2012, pp. 17-34, http://hal.inria.fr/hal-00718748

[69] B. BUCHER, F. LE BER. *Introduction*, in "Innovative Software Development in GIS", B. BUCHER, F. LE BER (editors), GIS Series, ISTE - WILEY, 2012, pp. 1-21, http://hal.inria.fr/hal-00724078

[70] F. LE BER, C. BRASSAC. *Modéliser "l'entre deux" dans l'organisation spatiale des exploitations agricoles – Mise en évidence de quelques problématiques*, in "Géoagronomie, paysage et projets de territoire", S. LARDON (editor), Indisciplines, QUAE, October 2012, pp. 63-72, http://hal.inria.fr/hal-00742683

[71] F. LE BER, B. BUCHER. *Analyse des spécificités des développements logiciels en géomatique*, in "Développements logiciels en géomatique – innovations et mutualisation", B. BUCHER, F. LE BER (editors), IGAT, Hermes Lavoisier, 2012, pp. 264-269, http://hal.inria.fr/hal-00718762

[72] F. LE BER, B. BUCHER. *Analysis of the specificities of Software Development in Geomatics Research*, in "Innovative Software Development in GIS", B. BUCHER, F. LE BER (editors), GIS Series, ISTE - WILEY, 2012, pp. 285-292, ISBN : 978-1848213647, http://hal.inria.fr/hal-00724081

[73] F. LE BER, J.-F. MARI. *GenExp-LandSiTes : un logiciel générateur de paysages agricoles 2D*, in "Développements logiciels en géomatique", B. BUCHER, F. LE BER (editors), Information Géographique et Aménagement du Territoire, HERMES Lavoisier, July 2012, pp. 181 – 202, http://hal.inria.fr/hal-00718266

[74] F. LE BER, J.-F. MARI. *GenExp-LandSiTes: a 2D Agricultural Generating Piece of Software*, in "Innovative Software Development in GIS", B. BUCHER, F. LE BER (editors), GIS, ISTE WILEY, July 2012, pp. 189 – 214, http://hal.inria.fr/hal-00718259

[75] D. RITCHIE. *Modeling Protein-Protein Interactions by Rigid-Body Docking*, in "Drug Design Strategies: Computational Techniques and Applications", T. CLARK, L. BANTING (editors), RSC Drug Discovery Series, RSC Publishing, 2012, pp. 56-86 [*DOI :* 10.1039/9781849733403], http://hal.inria.fr/hal-00666809

Books or Proceedings Editing

[76] S. O. KUZNETSOV, A. NAPOLI, S. RUDOLPH (editors). , *Proceedings of the ECAI Workshop on Formal Concept Analysis for Artificial Intelligence (FCA4AI)*, CEUR Proceedings, CEUR Proceedings (http://ceur-ws.org/Vol-939/)Montpellier, France, 2012, vol. 939, 88 p. , http://hal.inria.fr/hal-00768961

Research Reports

[77] V. CODOCEDO, I. LYKOURENTZOU, A. NAPOLI. , *Semantic Indexing and Retrieval based on Formal Concept Analysis*, Inria, June 2012, http://hal.inria.fr/hal-00713202

[78] E. EGHO, C. RAÏSSI, T. CALDERS, T. BOURQUARD, N. JAY, A. NAPOLI. , *On Measuring Similarity for Sequences of Itemsets*, Inria, October 2012, n^{o} RR-8086, 19 p. , http://hal.inria.fr/hal-00740231

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] O. BODENREIDER, R. STEVENS. *Bio-ontologies: current trends and future directions*, in "Briefings in Bioinformatics", 2006, vol. 7, n^{o} 3, pp. 256–274

[89] 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

[90] A. COULET, Y. GARTEN, M. DUMONTIER, R. B. ALTMAN, M. A. MUSEN, N. H. SHAH. *Integration and publication of heterogeneous text-mined relationships on the Semantic Web*, in "Journal of Biomedical Semantics", May 2011, vol. 2, n^{o} S2, S10 p. , http://hal.archives-ouvertes.fr/hal-00585215

[91] A. COULET, N. H. SHAH, Y. GARTEN, M. A. MUSEN, R. B. ALTMAN. *Using text to build semantic networks for pharmacogenomics*, in "Journal of Biomedical Informatics", Dec 2010, vol. 43, n^{o} 6, pp. 1009–1019 [*DOI :* 10.1016/J.JBI.2010.08.005], http://hal.inria.fr/inria-00549695

[92] 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, pp. D211–D222

[93] E. GAILLARD, J. LIEBER, E. NAUER. *Adaptation knowledge discovery for cooking using closed itemset extraction*, in "The Eighth International Conference on Concept Lattices and their Applications - CLA 2011", Nancy, France, October 2011, http://hal.inria.fr/hal-00646732/en

[94] 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

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

[96] 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

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

[98] 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

[99] C. JONQUET, P. LEPENDU, S. FALCONER, A. COULET, N. F. 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

[100] 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

[101] 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

[102] 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

[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] 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, pp. W445-W449 [*DOI :* 10.1093/NAR/GKQ311], http://hal.inria.fr/inria-00522712/en

[106] 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

[107] J.-F. MARI, F. LE BER. *Temporal and Spatial Data Mining with Second-Order Hidden Models*, in "Soft Computing", 2006, vol. 10, n° 5, pp. 406–414

[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] V. PÉREZ-NUENO, S. PETTERSSON, D. RITCHIE, J. BORRELL, J. TEIXIDÓ. *Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening*, in "Journal of chemical information and modeling", Apr 2009, vol. 49, n° 4, pp. 810-823 [*DOI :* 10.1021/CI800468Q], http://hal.inria.fr/inria-00434261/en

[110] V. PÉREZ-NUENO, D. RITCHIE. *Using Consensus-Shape Clustering To Identify Promiscuous Ligands and Protein Targets and To Choose the Right Query for Shape-Based Virtual Screening*, in "Journal of Chemical Information and Modeling", May 2011, vol. 51, n° 6, pp. 1233-1248 [*DOI :* 10.1021/CI100492R], http://hal.inria.fr/inria-00617922

[111] 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° 11, pp. 2146–2165

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

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

[114] D. RITCHIE, V. VENKATRAMAN. *Ultra-fast FFT protein docking on graphics processors*, in "Bioinformatics", 2010, vol. 26, n° 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, 2003 p.

[116] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *A proposal for combining Formal Concept Analysis and description Logics for mining relational data*, in "Proceedings of the 5th International Conference on Formal Concept Analysis (ICFCA 2007), Clermont-Ferrand", S. O. KUZNETSOV, S. SCHMIDT (editors), LNAI 4390, Springer, Berlin, 2007, pp. 51–65

[117] A. RUTTENBERG, T. CLARK, W. J. BUG, M. SAMWALD, O. BODENREIDER, H. CHEN, D. DOHERTY, K. FORSBERG, Y. GAO, V. KASHYAP, J. KINOSHITA, J. S. LUCIANO, M. S. MARSHALL, C. OGBUJI, J. REES, S. STEPHENS, G. T. WONG, E. WU, D. ZACCAGNINI, T. HONGSERMEIER, E. NEUMANN, I. HERMAN, K.-H. CHEUNG. *Advancing translational research with the Semantic Web*, in "BMC Bioinformatics", 2007, vol. 8, n^{o} S-3

[118] M. SAMWALD, A. COULET, I. HUERGA, R. L. POWERS, J. S. LUCIANO, R. R. FREIMUTH, F. WHIPPLE, E. PICHLER, E. PRUD'HOMMEAUX, M. DUMONTIER, M. S. MARSHALL. *Semantically enabling pharmacogenomic data for the realization of personalized medicine*, in "Pharmacogenomics", Jan 2012, vol. 13, n^{o} 2, pp. 201–212 [*DOI :* 10.2217/PGS.11.179], http://hal.inria.fr/hal-00752095

[119] A. STEIN, A. CEOL, P. ALOY. *3did: identification and classification of domain-based interactions of known three-dimensional structure*, in "Nucleic Acids Research", 2010, vol. 39, pp. D718–D723

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

[121] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN, A. BOC, V. MAKARENKOV. *Fast Mining of Iceberg Lattices: A Modular Approach Using Generators*, 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, October 2011, http://hal.inria.fr/hal-00640898/en

[122] 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

[123] 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

[124] 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

[125] 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

[126] 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

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