INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Orpailleur

Knowledge Discovery guided by Domain Knowledge

Nancy - Grand Est

Theme : Knowledge and Data Representation and Management

Activity Report

2010

Table of contents

Orpailleur is a project-team at INRIA Nancy-Grand Est and LORIA since the beginning of 2008. It is a rather large and special team as it includes computer scientists, but also a biologist, chemists, and a physician. Life sciences, chemistry, and medicine, are application domains of first importance and the team develops working systems for these domains.

1. Team

Research Scientists

Amedeo Napoli [Team leader, Researcher (DR CNRS), HdR]
Marie-Dominique Devignes [Researcher (CR CNRS), HdR]
Bernard Maigret [Researcher (DR CNRS), HdR]
Chedy Raïssi [Researcher (CR INRIA)]
Dave Ritchie [Chaire Excellence ANR, DR INRIA since October 2010]
Yannick Toussaint [Researcher (CR INRIA)]

Faculty Members

Adrien Coulet [Associate Professor (MdC ESIAL Université Henri Poincaré Nancy, from September 2010)]
Nicolas Jay [Associate Professor (Faculté de Médecine, UHP Nancy]
Florence Le Ber [Professor (ENGEES Strasbourg), HdR]
Bart Lamiroy [Associate Professor (délégation INRIA, MdC Institut National Polytechnique de Lorraine)]
Jean Lieber [Associate Professor (MdC Université Henri Poincaré Nancy 1), HdR]
Jean-François Mari [Professor (Université de Nancy 2), HdR]
Emmanuel Nauer [Associate Professor (MdC Université Paul Verlaine Metz)]
Malika Smaïl-Tabbone [Associate Professor (MdC Université Henri Poincaré Nancy 1)]

Technical Staff

Alexandre Blansché [Engineer (until August 2010)]
Inaki Fernandez [Engineer (from February 2010)]
Renaud Grisoni [Engineer (from October 2010)]
Jean-François Kneib [Engineer MBI (from November 2010)]
Luis Felipe Melo [Engineer (from November 2010)]
Birama NDiayé [Engineer]

PhD Students

Zainab Assaghir [ATER UHP, Thesis defended in November 2010]
Yasmine Assess [PhD Student (INCa Grant)]
Isiru Bayissa [PhD Student (BioIntelligence and Lorraine Region Grant, from October 2010)]
Sid-Ahmed Benabderrahmane [PhD Student (INCa Grant)]
Rahma Boujelbane [PhD Student (BioIntelligence Grant, from October 2010)]
Emmanuel Bresso [PhD Student (Cifre Harmonic Pharma)]
Julien Cojan [PhD Student (AMX Grant, ATER from October 2010)]
Sébastien Da Silva [PhD Student (INRA - INRIA Grant, from October 2010)]
Valmi Dufour-Lussier [PhD Student (MERT Grant, from October 2010)]
Elias Egho [PhD Student (ANR Trajcan Project, from October 2010)]
Leo Ghemtio [PhD Student (ANR Grant), Thesis defended in May 2010, Post-Doctoral fellow until December 2010 (CNRS Grant)]
Anisah Ghoorah [PhD Student (ANR Contract)]
Ana Karena [PhD Student (“co-tutelle” student, from May 2010 until February 2011)]
Mehdi Kaytoue [PhD Student (MERT Grant and ATER from October 2010)]
Thomas Meilender [PhD Student (CIFRE, A2ZI Company)]

Post-Doctoral Fellows

Thomas Bourquard [Post-Doctoral fellow (ANR Grant, from July 2010)]
Lazaros Mavridis [Post-Doctoral fellow (ANR Grant)]

Violeta Pérez-Nueno [Post-Doctoral fellow (ERC Marie Curie Grant, from July 2010)]
Lian Shi [Post-Doctoral fellow (Inria Grant, from February 2010)]
Vishwesh Venkatraman [Post-Doctoral fellow (ANR Grant)]
Jean Villerd [Post-Doctoral fellow (ANR Vigitermes Project and Lorraine Region Grant, until October 2010)]

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 its 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

The Taaable system [35] won the first prize and the adaptation challenge of the 2010 "Computer Cooking Contest" (CCC), at the 18th International Conference on Case-Based Reasoning (ICCBR), in Alessandria (Italy). This is the third year that the Taaable system is developed for participating in this challenge. The Taaable system is available on line at http://taaable.fr. This system is designed with the collaboration of the Score Team at LORIA and of the SILEX team (LIRIS Lyon). The Taaable system won the second prize in the first CCC [87] at the European Conference on Case-Based Reasoning in Trier (Germany, September 2008) The system also won the the second prize in the second "Computer Cooking Contest" (at ICCBR-2009, Seattle, USA) [88]. The design of the Taaable system involves a large part of the Orpailleur team, and needs joint efforts and combination of many skills and capabilities, such as knowledge representation, ontology engineering, classification, case-based reasoning, text-mining, information retrieval, and semantic wikis.

The application of KDDK process in the domain of Life Sciences made progress in 2010. Fast algorithms for 3D-shape classification and docking were designed, based on spherical harmonics and GPU programming. Dave Ritchie and others have organized a successful blind shape comparison experiment in the framework of Eurographics Workshop on 3D Object Retrieval (SHREC'10 Track:Protein Models, Norrköping, Sweden, 2010). The GPU version of the well-known Hex docking program is more than 45 times faster in docking computation (http://hexserver.loria.fr).

3. Scientific Foundations

3.1. From KDD to KDDK

Glossary

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 [101]) [112].
- Numerical methods are based on second-order Hidden Markov Models (HMM2, designed for pattern recognition [109]). 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") [108]. 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 [85], [86].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

Glossary

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 hierarchy [101]. Concept lattices are sometimes also called Galois lattices [89].

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 [57], [30].

From the numerical point of view, a Hidden Markov Model (HMM2) is a stochastic process aimed at extracting and modeling a stationary distribution of events. These models can be used for data mining purposes, especially for spatial and temporal data as they show good capabilities to locate stationary segments [110]). one special research effort focuses on the study of the application of HMM2 to composite data, both in the temporal and spatial domain, to produce a multi-dimensional classification based on multiple attributes.

3.3. Elements on Text Mining

Glossary

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 [84], [97]. 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 [92], [91], [90]. 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 [32]. 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

Glossary

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 [83]). 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 [36]. 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, Sid-Ahmed Benabderrahmane, Rahma Boujelbane, Thomas Bourquard, Emmanuel Bresso, Marie-Dominique Devignes, Elias Egho, Léo Ghemthio, Anisah Ghoorah, Renaud Grisoni, Nicolas Jay, Mehdi Kaytoue, Bernard Maigret, Lazaros Mavridis, Amedeo Napoli, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Yannick Toussaint, Vishwesh Venkatraman.

Glossary

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. The Kasimir Project

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

The Kasimir research project holds on decision support and knowledge management for the treatment of cancer [107]. 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. Currently, technologies of the semantic Wikis [81] are adapted for the management of decision protocols.

5. Software

5.1. Generic Symbolic KDD Systems

5.1.1. The Coron Platform

Participants: Mehdi Kaytoue [contact person], Amedeo Napoli, Yannick Toussaint.

The Coron platform [121] [80], [43] is a KDD toolkit organized around three main components: (i) Coron-base, (ii) AssRuleX, and (iii) pre- and post-processing modules. The software has been registered at the "Agence pour la Protection des Programmes" (APP) and is freely available[1]. 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].

This program implements the algorithms described in VLDB 2010's research paper "Computing Closed Skycubes" (major [9]). 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].

[1] http://coron.loria.fr

CarottAge [2] is a data mining system, freely available (GPL license) and based on Hidden Markov Models of second order. 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.

In these practical applications, the system aims at building a partition –called the hidden partition– in which the inherent noise of the data is withdrawn as much as possible. The CarottAge 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).

5.2.2. GenExp-LandSiTes: KDD and simulation

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

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 [47]. 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-INRIA project about landscape modeling, PAYOTE (2009-10), that gathers eleven research teams of agronomists, ecologists, statisticians, and computer scientists. The PAYOTE project is continuing in 2011, focusing on the comparison of various methods for analyzing and building temporal and spatial landscape structures. The PhD thesis of Sébastien da Silva is beginning within this framework and is conducted both by Claire Lavigne (DR in ecology, INRA Avignon) and Florence Le Ber.

5.3. KDD in Systems Biology

5.3.1. Automatic extraction of metadata for biological database retrieval and discovery (BioRegistry)

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

There are a growing number of biological databases which deal with the huge amount of data produced by genomic and post-genomic research. The need for a well-maintained indexed directory is therefore an important issue to make full use of these databases. The BioRegistry repository aims at associating content metadata with biological databases in view of retrieval or discovery. It is automatically generated from a publicly available list of biological databases (The Molecular Biology Database Collection published in Nucleic Acids Research). The content metadata are terms belonging to a biomedical thesaurus. Querying modalities have been implemented including a search by semantic similarity. A classification method based on extended FCA allows a user to browse and discover databases through the BioRegistry [17]. The BioRegistry repository is available at http://bioregistry.loria.fr.

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

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

MOdel-driven Data Integration for Mining (MODIM).

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

A position of engineer ("Ingénieur Jeune Diplomé INRIA") was granted to the Orpailleur team to develop the MODIM software (MOdel-driven Data Integration for Mining). This software for data integration can be summarized along three steps: (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 is undergoing an INRIA APP procedure on December, 2010.

5.4. Knowledge-Based Systems and Semantic Web Systems

5.4.1. CreChainDo

Participants: Emmanuel Nauer [contact person], Yannick Toussaint.

The "CreChainDo" system is aimed at information retrieval on the web and is based on FCA [33]. This system is inspired from the Credo system [94], [95]. Given a user query, the Credo system returns a list of documents retrieved by a search engine within a concept lattice allowing search and navigation. The CreChainDo system extends the Credo approach in introducing interaction and iteration in the design of concept lattices organizing documents returned by Google. The concept lattices help the user to explore the search results in a structured and synthetic way. The CreChainDo system is available on the web[3].

5.4.2. The Kasimir System for Decision Knowledge Management

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

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 [126]. Actually, the mining process in CabamakA is implemented thanks to a frequent close itemset extraction module of the Coron platform (see §5.1.1).

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

Participants: Alexandre Blansché, Julien Cojan, Valmi Dufour-Lussier, Jean Lieber [contact person], Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

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 CBR, ontology engineering, text mining, text annotation, knowledge representation, and hierarchical classification [35]. 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.

[3] http://intoweb.loria.fr/CreChainDo

The Taaable system is available on line at http://taaable.fr. In addition, Taaable won the second price in the first “Computer Cooking Contest” [87] (European Conference on Case-Based Reasoning, September 2008, Trier, Germany), and in the second “Computer Cooking Contest” (International Conference on Case-Based Reasoning, July 2009, Seattle, USA) [88]. In 2010, it won the first price and the adaptation challenge. Indeed, it has proposed two new adaptation approaches: adaptation of quantities and adaptation of recipe text preparations [39].

6. New Results

6.1. The Mining of Complex Data

Participants: Zainab Assaghir, Isiru Bayissa, Alexandre Blansché, Elias Egho, Nicolas Jay, Mehdi Kaytoue, Florence Le Ber, Luis Felipe Melo, Amedeo Napoli, Chedy Raïssi, Lian Shi, Yannick Toussaint, Jean Villerd.

Formal concept analysis, 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 ease of use, on the efficiency of the methods, and on the ability to fit evolving situations. Accordingly, the team is working on extensions of these symbolic methods to be applied on complex data such as objects with multi-valued attributes (e.g. domains or intervals), n-ary relations, sequences, trees, graphs, documents.

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 [118]. 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 [92], [91].

Another extension of FCA is based on Pattern Structures (PS) [100], which allows to build a concept lattice from complex data, e.g. nominal, numerical, and interval data. In (major [6]), 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 [68], [42], [67]. Moreover, similarity is not a transitive relation and this lead us to the study of tolerance relations [41]. In addition, a new research perspective is aimed at using frequent itemset search methods for mining interval-based data being guided by pattern structures (major [6]).

Pattern structures in association with a similarity measure were applied in the field of decision support in agronomy. In this domain, a set of agro-ecological indicators is aimed at helping farmers to improve their agricultural practices by estimating the impact of cultivation practices on the “agrosystem”. The modeling and the assessment of environmental risk require a large number of parameters whose measure is imprecise. The propagation of the imprecision and the different types of imprecision have to be taken into account in the computation of the value of indicators for decision support. Actually, based on pattern structures with a associated similarity measure, this problem has been approached as an information fusion problems with substantial results [60], [34], [59] [12].

Still in the context of agronomy, research work is in concern with the design of representation and reasoning models of spatial structures in knowledge-based systems, and in parallel, with the mining of complex hydrobiological data with concept lattices. FCA was compared and combined with statistical approaches to deal with multi-valued contexts in hydrobiology ([66], [69] major [2])

For completing the work on itemset search, there is still on-going work on frequent and rare itemset search for various reasons, among which improving standard algorithms, being able to build lattices from very large data, and completing the algorithm collection of the Coron platform. This year, substantial results were obtained on the search for rare itemsets which is an activity very important in biology and medicine because of the existence of rare symptoms [57], [30].

6.1.2. Privacy, anonymization, skylines, and streams

In the past decade, most of the research in privacy preserving data mining has been focusing on the privacy issues for relational data. Techniques such as k-anonymity, l-diversity and t-closeness have been proposed to address related problems. The publication of transaction data, such as market basket data, medical records, and query logs, serves the public benefit. Mining such data allows for the derivation of association rules that connect certain items to others with measurable confidence. Still, this type of data analysis poses a privacy threat; an adversary having partial information on a person's behavior may confidently associate that person to an item deemed to be sensitive. Ideally, an anonymization of such data should lead to an inference-proof version that prevents the association of individuals with sensitive items, otherwise allowing truthful associations to be derived. Original approaches to this problem were based on value perturbation, damaging data integrity. Recently, value generalization has been proposed as an alternative; still, approaches based on it have assumed either that all items are equally sensitive, or that some are sensitive and can be known to an adversary only by association, while others are non-sensitive and can be known directly. Yet in reality there is a distinction between sensitive and non-sensitive items, but an adversary may possess information on any of them. Most critically, no antecedent method aims at a clear inference-proof privacy guarantee. In our research work, we propose the first, to our knowledge, privacy concept that inherently safeguards against sensitive associations without constraining the nature of an adversary's knowledge and without falsifying data [37].

Recently, skyline analysis has attracted a lot of interest due to its importance in multi-criteria decision making applications. In our research work, we introduce a novel approach significantly reducing domination tests for a given subspace and the number of subspaces searched (major [9]). Technically, we identify two types of skyline points that can be directly derived without using any domination tests. Moreover, based on formal concept analysis, we introduce two closure operators that enable a concise representation of skyline cubes. We show that this concise representation is easy to compute and develop an efficient algorithm, which only needs to search a small portion of the huge search space.

Sampling streams of continuous data with limited memory, or "reservoir sampling", is a utility algorithm. Standard reservoir sampling maintains a random sample of the entire stream as it has arrived so far. This does not meet the requirement of many applications to give preference to recent data. The simplest algorithm for maintaining a random sample of a sliding window reproduces periodically the same sample design. This is undesirable for many applications. In our research work, we propose an effective algorithm, which is very simple and therefore very efficient, for maintaining -almost- a random sample of a sliding window [48].

6.1.3. KDDK in Text Mining

Ontologies help software and human agents to communicate by providing shared and common domain knowledge, and by supporting various tasks, e.g. problem-solving and information retrieval. In practice, building an ontology depends on a number of "ontological resources" having different types: thesaurus, dictionaries, texts, databases, and ontologies themselves. We are currently working on the design of a methodology and the implementation of a system for ontology engineering from heterogeneous ontological resources [32]. This methodology is based on both FCA and RCA, and was previously successfully applied in contexts such as astronomy and biology.

This year, an engineer will be in charge of implementing a new and robust system being guided by the previous research results and opening some new research directions involving trees and graphs.

Besides text mining, pharmacovigilance (PV) is in concern with the study and the prevention of adverse reactions to drugs (ADR), based on data collected by specialized centers and stored in case report databases (CRDBs). The CRDBs are then mined for finding unexpected associations between drugs and ADR that can be interpreted as signals. One objective of the ANR Project Vigitermes, which ended in June 2010, was to design a knowledge-based system for the management and the documentation of case reports, and, as well, for the detection of unexpected pharmacological associations. Following expert needs, we propose a method based on FCA for identifying candidates for pharmacological associations to be investigated in clinical trials (major [11]). In addition, this identification method uses statistical components for filtering significant associations. It was implemented within a prototype system and validated through an experiment on a database from the "Georges Pompidou" hospital.

Another work in text mining is concerned with the extraction of pharmacogenomics relationships from texts. A large amount of biomedical knowledge is lying in texts embedded in published articles, clinical files or biomedical public databases. For building operational knowledge bases from these textual sources, it is important to capture and formalize this knowledge. Here, relationships (also known as facts or events in the NLP literature) between biological entities represent elementary but interesting and reusable knowledge units.

In (major [4]), we propose a method based on a syntactic parsing for extracting rich semantic relationships between pairs of entities co-occurring in a single sentence. The method was applied in pharmacogenomics (study of the impact of individual genomic variation on drug responses) and we obtained a resource encoded in RDF that summarizes pharmacogenomics relationships mentioned into roughly 17 million Medline abstracts. This resource appears to be of major interest since it is used to guide human curation of biomedical databases, and to derive new knowledge about drug-drug interactions [102].

6.1.4. KDDK in Chemical Reaction databases

The mining of chemical chemical reaction databases is an important task for at least two reasons: (i) the challenge represented by this task regarding KDDK, (ii) the industrial needs that can be met whenever substantial results are obtained. Chemical reactions are complex data, that may be modeled as undirected labeled graphs. They are the main elements on which synthesis in organic chemistry relies, knowing that synthesis —and thus chemical reaction databases— is of first importance in chemistry, but also in biology, drug design, and pharmacology. From a problem-solving point of view, synthesis in organic chemistry must be considered at two main levels of abstraction: a strategic level where general synthesis methods are involved –a kind of meta-knowledge– and a tactic level where specific chemical reactions are applied. An objective for improving computer-based synthesis in organic chemistry is to discover general synthesis methods from currently available chemical reaction databases for designing generic and reusable synthesis plans. Graph-mining methods have been successfully used for the discovery of general synthesis methods in collaboration with chemists and in accordance with needs of chemical industry (major [8]).

6.2. KDDK in Life Sciences

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Emmanuel Bresso, Rahma Boujelbane, Thomas Bourquard, Adrien Coulet, Marie-Dominique Devignes, Léo Ghemthio, Anisah Ghoorah, Renaud Grisoni, Ana Karena, Mehdi Kaytoue, Jean-François Kneib, Florence Le Ber, Bernard Maigret, Jean-François Mari, Lazaros Mavridis, Amedeo Napoli, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Vishwesh Venkatraman.

One of the major challenges in the post genomic era consists in analyzing terabytes of biological data stored in hundreds of heterogeneous databases (DBs). The extraction of knowledge units from these large volumes of data would give sense to the present data production effort with respect to domains such as disease understanding, drug discovery, and pharmacogenomics or systems biology. Research reported here addresses these important issues and shows the spreading of KDDK over such domains.

6.2.1. Knowledge Discovery from Transcriptomic Data

This work is in concern with the application of classification approaches to interpret transcriptomic data from colorectal cancer samples. This year, we proposed a new measure called IntelliGO which computes semantic similarity between genes for discovering biological functions shared by genes showing the same expression profiles. This measure takes into account domain knowledge represented in Gene Ontology (GO). An original annotation vector space model is defined in which a dimension represents an annotation term and the dot product between two terms is calculated with the generalized cosine similarity measure that handles the fact that these terms may share semantic relationships, i.e. common ancestors in the GO graph. Moreover a weighting scheme includes the information content of each term with respect to a given annotation corpus as well as a customized value for each type of GO evidence code. The IntelliGO similarity measure was tested on two benchmarking datasets consisting of biological pathways (KEGG database) and functional domains (Pfam database, major [1]). An on-line version of the IntelliGO measure is available at http://bioinfo.loria.fr/Members/benabdsi/intelligo_project/.

6.2.2. Relational data mining applied to 3D protein patches for characterizing and predicting interaction sites

Protein-Protein Interactions (PPIs) play crucial roles in living systems and their understanding is important for the investigation of complex biological processes. Great effort has been put into both experimental and computational methods to identify or predict PPIs [119]. However, accurate prediction of PPI is still a challenging task because the available protein 3D structures are under exploited and PPI sites are not characterized explicitly. In this context, we have developed a method that aims at exploiting the set of available 3D structures in PPI prediction, trying to push forward the limitations of the current approaches (qualified as black-boxes). Firstly we propose a relational representation of protein 3D patches. These patches correspond either to positive or negative examples of PPI sites. Then a relational data mining method, based on Inductive Logic Programming (ILP), is applied on the descriptions of 3D patches in order to learn by induction a general definition of the protein interaction site concept. At the moment, this work was presented as a poster at the JOBIM Conference in Montpellier (September, 2010), and it was presented during the life science day of the Charles Hermite Federation in Nancy (October 2010). Publications on this theme are in preparation.

6.2.3. A KDD approach for designing filters to improve virtual screening

In silico screening methodologies are widely recognized as efficient approaches in early steps of drug discovery. However, in the virtual high-throughput screening (VHTS) context, where hit compounds are searched among millions of candidates, three-dimensional comparison techniques and knowledge discovery from databases should offer a better efficiency to finding novel drug leads than those of computationally expensive molecular dockings. Therefore, we aimed to develop a filtering methodology that efficiently eliminates unsuitable compounds in VHTS process. Several filters were evaluated. The two first filters are structure-based and rely on either geometrical docking or pharmacophore depiction. The third filter is ligand-based and uses knowledge-based and fingerprint similarity techniques. These filtering methods have been tested with the Liver X Receptor (LXR), a target of therapeutic interest as LXR is a key regulator in maintaining cholesterol homeostasis. The results show that the three considered filters are complementary so that their combination should generate consistent compound lists of potential hits (major [5]).

6.2.4. Mining Agronomical and Biological Data with HMMs

Thanks to the CarottAge data mining system, we have carried out a study on the Niort plain (West of France) database. On this database, provided by the CEBC (UPR CNRS), the land use occupations of the fields covering a $400km^2$ area have been being recorded during 12 years. Based on farm surveys, we were able to retrieve and quantify changes in land use occupation. Besides, stochastic regularities –like the changes in the neighborhood system at the crop succession level– revealed by the stochastic modeling were proposed for interpretation by the analyst. These regularities were explained by individual farmer decision rules thanks to farm surveys [82]. Finally, the results of our analysis can be reused for modeling nitrate flow and for evaluating water pollution risks in a watershed.

In the biological field, we have implemented a new data mining method based on second-order HMM and combinatorial methods for Sigma Factor Binding Site (SFBS) prediction and Horizontal Gene Transfer (HGT) [99] detection that voluntarily implements a minimum amount of knowledge[4]. The original features of the

[4] C. Eng, A. Thibessard, M. Danielsen, T. Rasmussen, J.-F. Mari, and P. Leblond, In silico prediction of horizontal gene transfer in Streptococcus thermophilus, accepted for publication in Archives of Microbiology, 2011.

presented methodology include (i) the use of the CarottAge framework, (ii) an automatic area extraction algorithm that captures atypical DNA motifs of various size based on the variation of the state a posteriori probability, and (iii) a set of post processing algorithms suitable to the biologic interpretation of these segments. Two different HMM (M2-M0 HMM and M2-M2 HMM) have extracted meaningful regularities that are of interest in the area of promoter and HGT detection. The additional dependencies implemented in the M2-M2 HMM smooth dramatically the a posteriori probability. This smoothing effect allows the extraction of wider regularities in the genome as it has been shown in the HGT application.

When using CarottAge, the extraction of regularities in all these applications was achieved following the same mining scenario that starts by the estimation of a linear HMM to get initial seeds for the probabilities and, next, a linear to ergodic transform followed by a new estimation by the forward backward algorithm. Even if the data do not suit the model, the HMM can give interesting results allowing the domain specialist to put forward some new hypothesis. Also, we have noticed that the data preparation is a time consuming process that conditions all further steps of the data mining process. Several ways of encoding elementary observations have been tried in all applications during our interactions with the domain specialists.

6.3. Structural Systems Biology and Docking

Participants: Thomas Bourquard, Marie-Dominique Devignes, Anisah Ghoorah, Bernard Maigret, Lazaros Mavridis, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Vishwesh Venkatraman.

The HPASSB project started in January 2009 following Dave Ritchie's successful application for funding to the ANR Chaires d'Excellence 2008 (Senior Courte Durée) programme. The overall aim of HPASSB is to help the building of a new Centre of Excellence in France in the emerging discipline of structural systems biology. The HPASSB project complements existing competencies in the Orpailleur team represented by M.-D. Devignes (CR CNRS) who is coordinating the MBI project (Modelling Biomolecules and their Interactions, http://bioinfo.loria.fr), Malika Smaïl-Tabbone (MCU Nancy University) who is working on data integration and relational data-mining approaches, and Bernard Maigret (DR CNRS) who has an extensive experience of molecular dynamics and virtual screening. 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. In October 2010, Dave Ritchie joined the permanent staff at INRIA Nancy. This will ensure that the activities of this project will continue well beyond the duration of the original ANR grant.

6.3.1. Accelerating protein docking calculations using graphics processors

In this framework, we have recently adapted the *Hex* protein docking software to use modern graphics processors (GPUs) to carry out the expensive FFT part of a docking calculation (major [10]). Compared to using a single conventional central processor (CPU), a high-end GPU gives a speed-up of 45 or more. Furthermore, the *Hex* code has been re-written to use multi-threading techniques in order to distribute the calculation over as many GPUs and CPUs as are available. Thus, a calculation which formerly took many minutes or several hours can now be performed in a matter of seconds on a modern desk-top computer. This advance will facilitate future docking-based studies of large-scale protein interaction networks and building multi-component molecular structures [53]. This software is publicly available at http://hex.loria.fr. A public GPU-powered server has also been created (http://hexserver.loria.fr) [24].

6.3.2. Eigen-Hex: Modeling protein flexibility during docking

Although the *Hex* protein docking software can often make reasonably good predictions about how two proteins might fit together, a major limitation of many current algorithms, including *Hex*, is that that they assume that proteins are rigid objects. In fact, proteins can be highly flexible, and the internal conformations of their atoms often change on going from the unbound forms in the free proteins to the bound conformations in the complex. We are developing a novel approach to model such flexibility using a principal component analysis (PCA) technique to identify and predict the main atomic motions during a docking calculation. Compared to rigid body docking, the results obtained so far are promising. This work was recently presented at the 3D Special Interest Group (3D-SIG) meeting of the 18th International Conference on Intelligent Systems for Molecular Biology.

6.3.3. 3D-Blast: A new approach for protein structure alignment and clustering

We have recently developed a new sequence-independent protein structure alignment approach called 3D-Blast, which is exploits the spherical polar Fourier (SPF) correlation technique used in the *Hex* protein docking software [117]. The utility of this approach has been demonstrated by clustering subsets of the CATH protein structure classification database [113] for each of the four main CATH fold types, and by searching the entire CATH database of some 12,000 structures using several protein structures as queries. Overall, the automatic SPF clustering approach agrees very well with the expert-curated CATH classification, and ROC-plot analyses of database searches show that the approach has very high precision and recall. Database query times can be reduced considerably by using a simple rotationally-invariant pre-filter in tandem with a more sensitive rotational search with little or no reduction in accuracy. Hence it should soon be possible to perform on-line 3D structural searches in interactive time-scales [49]. This approach recently performed very well in a blind shape comparison experiment organised by Orpailleur as part of Eurographics Workshop on 3D Object Retrieval [50].

6.3.4. KDD-Dock: Protein docking using Knowledge-Based approaches

Protein docking is the difficult computational task of predicting how a pair of three-dimensional protein structures come together to form a complex. There is considerable interest in developing improved *ab initio* techniques which can make protein-protein docking predictions using only knowledge of their three-dimensional structures. The *Hex* docking program developed by Dave Ritchie is one such example. However, as structural genomics initiatives continue to populate the space of protein 3D structures, and as several on-line databases of protein interactions have recently become available, using structural database systems to perform docking by homology will become an increasingly powerful approach to predicting protein interactions. We recently used the SCOPPI [124] and 3DID [120] protein interaction databases to help make some very good predictions to two of the recent CAPRI target complexes, and we are now working to incorporate additional knowledge from other databases and to automate the overall approach. This work has been presented as a poster at the 9th European Conference on Computational Biology in Ghent, and was presented orally in Nancy at the Life Sciences theme day of the Charles Hermite Federation.

6.3.5. V-Dock: scoring protein-protein interactions using Voronoi fingerprints

There is growing interest in using computational docking techniques to help populate protein-protein interaction (PPI) networks [111]. Furthermore, recent computational cross-docking experiments indicate that the profile of scores for each pair-wise docking can reveal information about whether or not two proteins might associate, even if the precise binding mode cannot be calculated exactly [125]. The aim of this project is to investigate the use of Voronoi fingerprints of protein-protein interfaces [93] as a way to distinguish cognate and non-cognate pairs of proteins in large-scale cross-docking experiments. So far, the results seem very promising, and the approach was recently presented as a poster at the LIX Bioinformatics Colloquium at the Ecole Polytechnique.

6.3.6. 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 [115]. The main disease target in this project is the acquired immune deficiency syndrome (AIDS), caused by the human immuno-deficiency virus (HIV) [114]. However, the approach will be quite generic and will be broadly applicable to many other diseases. 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 protein sub-sites or which may bind multiple protein targets, and it will help to find new HIV entry-blocking compounds that target the CXCR4 and CCR5 cell surface receptor proteins. Recent progress on this project has been presented orally at the 5th Joint Sheffield Conference on Chemoinformatics and at the 6th German Conference on Chemoinformatics. The work was also presented at the 18th EuroQSAR Conference in Rhodes, and a comprehensive comparison and review of several current virtual screening algorithms has recently been published [31].

6.3.7. Critical Assessment of Protein-Protein Interactions (CAPRI challenge)

Every two years, the state of the art in protein docking is assessed at the CAPRI (Critical Assessment of PRedicted Interactions http://www.ebi.ac.uk/msd-srv/capri/) international conference. For the CAPRI assessments, participants are given the structures or sequences of a pair of proteins that are known to bind but for which the corresponding structure of the complex has not yet been published. Our method combines human expertise, fast rigid-body docking (using *Hex* program) and molecular dynamics. This strategy produced good results for several targets (our predictions were amongst the best for Target 34 according to the ligand RMSD criteria, and thanks to the literature and analogous interactions we identified the key interaction residues for Target 40 before they were revealed by the organisers). Overall, our approach correctly predicted three targets in the recently assessed rounds of CAPRI [106].

6.4. Around the Kasimir research project

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

This special research project involves researchers working around the Kasimir project and Bart Lamiroy who was attached to the Orpailleur Team during his “INRIA délégation” (2010) and at the same time was a visiting scientist at Lehigh University, USA. The background of Bart Lamiroy is in document and image analysis. Recently he was interested in investigating the application of KDDK to numerical and structural data including document images. The objective is to extend mining tools towards complex and semi-structured multi-media data on the one hand, and to associate image analysis with KDDK techniques on the other hand.

The main research direction which is followed at the moment is in concern with the Kasimir project. Actually, oncology protocols are mainly documented and represented in diagram formats. The classification and CBR techniques used in the Kasimir project require that the ontologies and decision protocols have to be represented in OWL. Based on previous work [105], [104], we started modeling the mapping of visual features in diagram charts with semantics of the medical domain ontology. The mapping between the visual ontology and the domain ontology should guide a more complete extraction of the protocols from the diagrams for completing the domain ontology of the Kasimir system.

Moreover, during his stay at Lehigh University, B. Lamiroy developed a new approach for recovering useful information within image data [45]. By recording a wide range of “provenance information” related to complex image analysis processes, the DAE platform (http://dae.cse.lehigh.edu) provides a large set of metadata that can be used by KDDK methods. For example, this allows the correlation and combination of numerical and symbolic aspects, e.g. relating image aspects and domain symbolic representations (within domain ontologies). This work bridges the gap between formal knowledge representation and signal-based pattern recognition and offers a robust experimental environment for further application of KDDK on image data.

6.5. Around the Taaable research project

Participants: Alexandre Blansché, Julien Cojan, Valmi Dufour-Lussier, Inaki Fernandez, Florence Le Ber, Jean Lieber, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint.

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 800 in 2008 and about 1500 in 2009 and in 2010) prevents from a manual indexing of recipes: this indexing is performed using semi-automatic techniques.

The first version of the Taaable system (2008) was the European vice-champion of the contest. The second version (2009) was the World vice-champion of the contest. The third version (2010) was the World champion: it has won the main challenge and the adaptation challenge [35]. A fourth version for the 2011's contest is under conception.

The partners of the 2010's Taaable project are members of Orpailleur and of Score (INRIA projects in Nancy). 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.

A general description of the 2010's Taaable system can be found in [35]. The most important original features of this version are:

A module for adapting quantities. In the previous versions of Taaable, only a substitution of ingredient types by other ingredient types was proposed by the system. Now, there is the possibility to adapt the ingredient quantities. In this way, there is a maximum preservation of some features of the global recipe, such as the quantity of sugar, of calories, etc. This implementation is based on a theoretical research published in 2009 [98].

A module for adapting recipe preparation texts. Another adaptation that was not studied before this year is the adaptation of the texts that describe the preparations [39]. Such an adaptation module has been implemented, using natural language processing techniques in order to transform recipes in a tree structure whose root is the final dish, whose leaves are the ingredients, and whose internal nodes represent the actions. The adaptation is performed on the tree structure and, thanks to links between the text and the tree, this adaptation has repercussions on the text.

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

- The representation of preparations in a temporal qualitative algebra [70].
- An algorithm for adapting cases defined in an expressive description logic (major [3] [62]).
- The study of the relations between rule-based adaptation and adaptation based on belief revision, that enables to incorporate rules in a revision-based adaptation [61].
- The study of the extension of the domain ontology to make the retrieval step of a case-based reasoning system more accurate [64], [63].

The fourth aspect involves text mining within CBR. In the Taaable system, similar cases are searched according to an ontology which is used to progressively refine or generalize a given target problem. The extension the domain ontology is based on the application of FCA on specific resources collected for this purpose. For example, for refining the ingredient hierarchy, a set of actions applied to ingredients are extracted from the text of recipes. The linguistic anaphoras in recipes require the use of a syntactic and dynamic semantic analysis for building a formal representation of a recipe [39] from which relations between ingredients and actions are extracted. Based on this textual analysis, the formal representation of the recipe can be considered as a tree structure whose root is the desired meal, whose leaves are ingredients, and whose internal nodes correspond to actions. In this way, a textual adaptation process can be defined where adaptation consists in a subtree substitution, i.e. replacing an initial subtree with a final and more accurate subtree. The search of the final subtree is based on FCA which is used to organize recipes w.r.t. their content.

7. Contracts and Grants with Industry

7.1. The BioIntelligence Project

Participants: Isiru Wakwoya, Rahma Boujelbane, Adrien Coulet, Marie-Dominique Devignes, Mehdi Kaytoue, Luis Felipe Melo, Amedeo Napoli [contact person], Chedy Raïssi, Malika Smaïl-Tabbone, Yannick Toussaint.

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).

Moreover, the “BioIntelligence” project is aimed at designing software modules for helping the biological daily practice and to guide knowledge discovery, knowledge representation and management, and finally innovation and production. 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. The kickoff meeting took place in Sophia-Antipolis between July 5th and 6th 2010.

Two thesis related to “BioIntelligence” are beginning in the Orpailleur team. The first one is in concern with ontology engineering for biology. Two main aspects which are considered at the moment are ontology matching and ontology mining for the design of ontology design patterns (involving graph mining methods). Among web resources, ontologies take a special place as they materialize models of the real-world, they provide the representation of domain knowledge, and they allow domain knowledge manipulation and dissemination. One objective of the thesis is to study and to design a complete process for discovering in heterogeneous resources domain models representing significant parts of these resources. Such models are also called ontology design patterns and are intended to be used as building blocks that can be interconnected for building a working domain ontology, based on a set of reusable components. Ontologies lying at the NCBO BioPortal (http://bioportal.bioontology.org/) will be considered as a training set for the thesis work.

The second 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.

7.2. The Spinal Image Project

Participants: Inaki Fernandez, Amedeo Napoli, Emmanuel Nauer [contact person].

This research work is based on an industrial collaboration with an enterprise called “The Picture Factory” whose commercial activity is to propose video documents (actually video rushes) on the web (http://www.thepicturefactory.fr/). The enterprise is using a system for managing the documents based on a database with three main components: a module for organizing, recording and accessing data (documents), a module for editing and indexing data, and a web interface allowing interactions with the system.

One objective of this project is to improve the management system in allowing an intelligent access and manipulation of the documents based on the content of these documents. Firstly, documents should be described and organized w.r.t. different points of view related to their content. Domain ontologies will be used for allowing such organization, and also for guiding the annotation of the documents. A special attention will be put on the relations between modeling, organization, annotation, and manipulation of the documents. Actually, knowledge editing and annotation will be performed within a semantic wiki. In this way, a dynamic evaluation of the performances of the system w.r.t. some intelligent tasks, e.g. search and access, will be achieved on-line for controlling the progressive improvements of the management system. Finally, a generalization to more classical search systems will be studied.

8. Other Grants and Activities

8.1. International projects and collaborations

8.1.1. Fapemig INRIA Project: Incorporating knowledge models into scalable data mining algorithms

Participants: Mehdi Kaytoue, Amedeo Napoli [contact person], Chedy Raïssi, Yannick Toussaint.

This Fapemig – INRIA research project 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.

In summary, the overall goal of this project is to enhance data mining algorithms targeted at text mining and graph mining by exploiting knowledge models that improve their effectiveness and by parallelizing them, so that they scale better.

8.1.2. International Collaborations in Biology and Chemistry

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Thomas Bourquard, Emmanuel Bresso, Marie-Dominique Devignes, Léo Ghemthio, Anisah Ghoorah, Bernard Maigret, Lazaros Mavridis, Violeta Pérez-Nueno, Dave Ritchie, Malika Smaïl-Tabbone, Vishwesh Venkatraman.

8.1.2.1. Grand Challenge project - Foundation Bill and Melinda Gates

This collaboration involves the “J. Craig Venter Institute” at Rockville, MD 20850 USA, and the “Centre International de Référence Chantal Biya pour la Recherche sur la Prevention et la Prise en charge du VIH/Sida” (CIRCB), BP 3077, Yaoundé Cameroun. It is entitled “Design and Setting Up of a Bioinformatics Platform Dedicated to HIV Drug Resistance Problems”.

An application for the Phase II funding program of the Grand Challenge Project was submitted in November 2010. This research work is currently under development and publications are in preparation.

8.1.2.2. Search for anti-HIV drugs acting as entry-blockers

In collaboration with computational chemistry colleagues at the University of Bari and the Institut Chimique de Saria (IQS) in Barcelona, Dave Ritchie has published reviews of the state of *in silico* protein structure modeling and *virtual drug screening* techniques for the CCR5 [96], and CXCR4 [26], entry-blocking molecules. As there now exist several hundred such entry-blockers, there is considerable interest in the chemoinformatics community in how best to use knowledge of known drug molecules to develop new and more potent new drug candidates [116]. The spherical harmonic clustering approach developed by Dave Ritchie and Violeta Pérez-Nueno was recently used successfully in a virtual screening study at the IQS to discover new high-affinity ligands for CXCR4 [114].

8.1.3. Mining complex data: International collaborations with UQAM Montréal and HSE Moscow

Participants: Mehdi Kaytoue, Amedeo Napoli, Chedy Raïssi, Yannick Toussaint.

Two close international research collaborations have still to be mentioned, which are based on scientific visits laboratories and the writing of common scientific papers.

- The first collaboration involves "Université du Québec à Montréal" (UQAM) in Montréal with Prof. Petko Valtchev and Laboratoire LIRMM in Montpellier with Prof. Marianne Huchard. Amedeo Napoli visited several times LIRMM in Autumn 2010, and UQAM Montréal in October 2010. The research topics are concerned with the design of algorithms for itemset search and association rule extraction, and the design of large concept lattices and relational concept lattices [65], [54], [57], [30].
- The second collaboration involves Sergei Kusnetsov at Higher School of Economics in Moscow (HSE). Mehdi Kaytoue and Amedeo Napoli visited HSE laboratory in July 2010 granted by the Poncelet Laboratory in Moscow, a joint CNRS – INRIA laboratory. Meanwhile Sergei Kuznetsov visited Loria in August and October 2010. The research topic in (ii) holds on extension of FCA algorithms for taking into account complex data (major [6], [41], [78], [79]).

8.2. National initiatives

8.2.1. ANR Nutrivigène

Participants: Mehdi Kaytoue, Amedeo Napoli [contact person], Jean Villerd.

Nutrigenomics is an emerging topic interested in elaborating dietary recommendations and new food products. In this context, "homocysteine" is an intermediate product of the carbon metabolism related to the status of folate and vitamin B12. Moreover, homocysteine is correlated with age and with vascular, cognitive, and neurological dysfunctions. Accordingly, the objective of the Nutrivigène project is to study whether, at the cellular level, "hyperhomocysteinemia" produces epigenetic changes of the expression of genes potentially related with the vascular, cognitive and neurological dysfunctions of volunteers of the cohort OASI which includes people recruited in a rural region of Sicily.

The Nutrivigène project involves various partners: INSERM U724 (Nancy Hospital) in association with IRCCS of Troina (Italy), INRA Alimentation Humaine (Clermont-Ferrand Theix), UMR CNRS 2738 (Marseille), LSGA (INPL Nancy), Nestlé-Waters (Vittel), and project-team Orpailleur (INRIA Nancy Grand Est). The role of Orpailleur Team is to apply KDD methods for analyzing the data in the OASI cohort and for evaluating the association between genetic data, homocysteine, and the vascular and cognitive functions of individuals. A set of experiments was carried out and results are currently analyzed for publication purposes.

8.2.2. 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, sex, place of residence. One thesis is beginning on this subject whose objective is to design a knowledge discovery system working on multidimensional and sequential data for characterizing patient care trajectories. This thesis will combine knowledge discovery and knowledge representation methods for improving the definition of patient care trajectories as temporal objects (sequential data mining). This research work will be also interested in decision support for improving healthcare in detecting for example typical or exceptional trajectories for planning with precision healthcare for a given population. This thesis work will extend a former research work based on Formal Concept Analysis and aimed at assisting domain experts for discovering groups of patients showing similar health condition, treatments or journeys through the healthcare system. FCA shows some capabilities for dealing with large amounts of data and for filtering (with a measure such as stability) very interesting results [103].

8.2.3. ANR Vigitermes: Data Mining for Pharmacovigilance

Participants: Yannick Toussaint, Jean Villerd.

Pharmacovigilance covers research activities related to detection, analysis, and prevention of unexpected adverse drug reactions (ADR). In France, ADRs when they are known have to be declared by health-care professionals. Besides, the regional "Pharmacovigilance Centers" collect spontaneous reports on ADRs for all drugs commercialized in France, while pharmacovigilance units of pharmaceutical laboratories receive spontaneous reports on ADRs in which they are directly concerned. All reports are registered in the pharmacovigilance national database, called AFSSaPS for "Agence Française de Sécurité Sanitaire des Produits de Santé". In the same way, at the international level, individual reports on unsuspected ADRs are collected and stored in a centralized database, including more than 3.7 million case reports, described in several languages.

The general objective of the Vigitermes project (which involves 10 partners) consists in supporting the work of Pharmacovigilance experts in two ways: firstly in guiding information retrieval and access to available resources, e.g. Pharmacovigilance database, product catalogs, medical literature, secondly, in improving signal detection in pharmacovigilance. In both cases, KDDK methods based on domain knowledge are used data analysis and knowledge discovery (major [58]).

8.3. Local initiatives

8.3.1. 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 M.-D. 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 has to be understood as “Automatic Processing of Languages and Knowledge” and 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.

8.3.2. Other initiatives

8.3.2.1. Cancéropole Grand-Est.

A collaboration with the “Laboratoire de Bioinformatique et Génomique Intégratives (LBGI)” at IGBMC Strasbourg involves a thesis funded by INCa (“Institut National du Cancer”) with a bipartite direction. This thesis is considered as one research operation within the annual meeting of “Canceropole Grand-Est”.

8.3.2.2. 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 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 pharmacologiques et effets secondaires de principes actifs”.

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, PKDD, ICFCA ...), as members of editorial boards, and finally in the organization of journal special issues.

  This year, Florence Le Ber was involved in the edition of three books and journal special number [76], [74], [77].

9.2. Teaching

- The members of the Orpailleur team are involved in teaching at all levels of teaching in the universities of Nancy, especially in Nancy Université including “Université Henri Poincaré Nancy-1”, “Université de Nancy-2”, “Institut Polytechnique de Lorraine”. Actually, most of the members of the Orpailleur team are employed on university positions.

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

10. Bibliography

Major publications by the team in recent years

[1] 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 [*DOI :* 10.1186/1471-2105-11-588], http://www.biomedcentral.com/1471-2105/11/588/abstract, http://hal.inria.fr/inria-00543910/en.

[2] A. BERTAUX, F. LE BER, A. BRAUD, M. TRÉMOLIÈRES. *Mining Complex Hydrobiological Data with Galois Lattices*, in "International Journal of Computing & Information Sciences", 2010, vol. 7, n^{o} 2, p. 63–77, http://hal.inria.fr/hal-00531756/en.

[3] J. COJAN, J. LIEBER. *An Algorithm for Adapting Cases Represented in an Expressive Description Logic*, in "18th International Conference on Case-Based Reasoning - ICCBR 2010 Case-Based Reasoning Research and Development", Alessandria Italie, I. BICHINDARITZ, S. MONTANI (editors), Lecture Notes in Artificial Intelligence, Springer Berlin, 07 2010, vol. 6176, p. 51-65, http://hal.inria.fr/inria-00506078/en/.

[4] A. COULET, N. SHAH, Y. GARTEN, M. MUSEN, R. ALTMAN. *Using text to build semantic networks for pharmacogenomics*, in "Journal of Biomedical Informatics", 2010, vol. 43, n^{o} 6, p. 1009–1019.

[5] L. GHEMTIO, M.-D. DEVIGNES, M. SMAÏL-TABBONE, M. SOUCHET, V. LEROUX, B. MAIGRET. *Comparison of three preprocessing filters efficiency in virtual screening: identification of new putative LXRbeta regulators as a test case*, in "Journal of chemical information and modeling", May 2010, vol. 50, n^{o} 5, p. 701–715, http://hal.inria.fr/hal-00547968/en.

[6] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI, S. DUPLESSIS. *Mining gene expression data with pattern structures in formal concept analysis*, in "Information Sciences", 2010, Article in press, http://hal.inria.fr/hal-00541100/en.

[7] E. G. LAZRAK, M. BENOÎT, J.-F. MARI. *Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales*, in "International Conference on Integrative Landscape Modelling", France Montpellier, Symposcience, February 2010, http://hal.inria.fr/inria-00482890/en.

[8] F. PENNERATH, G. NIEL, P. VISMARA, P. JAUFFRET, C. LAUREN, A. NAPOLI. *Graph-Mining Algorithm for the Evaluation of Bond Formability*, in "Journal of chemical information and modeling", January 2010, vol. 50, n^{o} 2, p. 221–239 [*DOI :* 10.1021/CI9003909], http://hal.inria.fr/hal-00471405/en.

[9] C. RAÏSSI, J. PEI, T. KISTER. *Computing Closed Skycubes*, in "Proceedings of the VLDB Endowment (PVLDB)", 2010, vol. 3, n^{o} 1, p. 838–847.

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

[11] J. Villerd, Y. Toussaint, A. L.-L. Louët. *Adverse Drug Reaction Mining in Pharmacovigilance Data Using Formal Concept Analysis*, in "Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, 2010, Proceedings,", J. L. Balcázar, F. Bonchi, A. Gionis, M. Sebag (editors), Lecture Notes in Computer Science 6323, Springer, 2010, p. 386–401.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[12] Z. Assaghir. *Analyse formelle de concepts et fusion de données. Application à l'estimation et au contrôle de l'incertitude des indicateurs agri-environnementaux*, Nancy Université / Institut National Polytechnique de Lorraine, France, November 2010.

[13] L. Ghemtio. *Simulation numérique et approche orientée connaissance pour la découverte de nouvelles molécules thérapeutiques*, Nancy Université / Université henri Poincaré, France, May 2010.

Articles in International Peer-Reviewed Journal

[14] 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 [*DOI :* 10.1186/1471-2105-11-588], http://www.biomedcentral.com/1471-2105/11/588/abstract, http://hal.inria.fr/inria-00543910/en.

[15] A. Bertaux, F. Le Ber, A. Braud, M. Trémolières. *Mining Complex Hydrobiological Data with Galois Lattices*, in "International Journal of Computing & Information Sciences", 2010, vol. 7, n^{o} 2, p. 63–77, http://hal.inria.fr/hal-00531756/en.

[16] A. Coulet, N. Shah, Y. Garten, M. Musen, R. Altman. *Using text to build semantic networks for pharmacogenomics*, in "Journal of Biomedical Informatics", 2010, vol. 43, n^{o} 6, p. 1009-19.

[17] M.-D. Devignes, P. Franiatte, N. Messai, E. Bresso, A. Napoli, M. Smaïl-Tabbone. *BioRegistry: Automatic extraction of metadata for biological database retrieval and discovery*, in "International Journal of Metadata Semantics and Ontologies", 2010, vol. 5, n^{o} 3, p. 184–193 [*DOI :* 10.1504/IJMSO.2010.034043], http://hal.inria.fr/inria-00502297/en.

[18] L. Ghemtio, M.-D. Devignes, M. Smaïl-Tabbone, M. Souchet, V. Leroux, B. Maigret. *Comparison of three preprocessing filters efficiency in virtual screening: identification of new putative LXRbeta regulators as a test case*, in "Journal of chemical information and modeling", May 2010, vol. 50, n^{o} 5, p. 701–715, http://hal.inria.fr/hal-00547968/en.

[19] L. Ghemtio, E. Jeannot, B. Maigret. *Efficiency of a hierarchical protocol for highthroughput structure-based virtual screening on Grid5000 cluster grid*, in "Open Access Bioinformatics", May 2010, vol. 2, p. 41–53 [*DOI :* 10.2147/OAB.S7272], http://hal.inria.fr/hal-00547970/en.

[20] X. Iturrioz, R. Alvear-Perez, N. De Mota, C. Franchet, F. Guillier, V. Leroux, H. Dabire, M. Le Jouan, H. Chabane, R. Gerbier, D. Bonnet, A. Berdeaux, B. Maigret, J.-L. Galzi, M. Hibert, C. Llorens-Cortes. *Identification and pharmacological properties of E339-3D6, the first nonpeptidic apelin receptor agonist*, in "The FASEB Journal", May 2010, vol. 24, n^{o} 5, p. 1506–1517, http://hal.inria.fr/hal-00547967/en.

[21] X. Iturrioz, R. Gerbier, V. Leroux, R. Alvear-Perez, B. Maigret, C. Llorens-Cortes. *By interacting with the C-terminal Phe of apelin, Phe255 and Trp259 in helix VI of the apelin receptor are critical for internalization.*, in "The Journal of Biological Chemistry", October 2010, vol. 285, n^{o} 42, p. 32627-32637, http://hal.inria.fr/hal-00547969/en.

[22] M. Kaytoue, S. O. Kuznetsov, A. Napoli, S. Duplessis. *Mining gene expression data with pattern structures in formal concept analysis*, in "Information Sciences", 2010, Article in press, http://hal.inria.fr/hal-00541100/en.

[23] E. G. Lazrak, J.-F. Mari, M. Benoît. *Landscape regularity modelling for environmental challenges in agriculture*, in "Landscape Ecology", Sept. 2010, vol. 25, n^{o} 2, p. 169 – 183.

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

[25] F. Pennerath, G. Niel, P. Vismara, P. Jauffret, C. Lauren, A. Napoli. *Graph-Mining Algorithm for the Evaluation of Bond Formability*, in "Journal of chemical information and modeling", January 2010, vol. 50, n^{o} 2, p. 221–239 [*DOI :* 10.1021/CI9003909], http://hal.inria.fr/hal-00471405/en.

[26] V. Pérez-Nueno, D. Ritchie. *Applying in silico Tools to the Discovery of Novel CXCR4 Inhibitors*, in "Drug Development Research", 12 2010, http://hal.inria.fr/inria-00550645/en/.

[27] V. Pérez-Nueno, V. Venkatraman, L. Mavridis, T. Clark, D. Ritchie. *Using spherical harmonic surface property representations for ligand-based virtual screening*, in "Molecular Informatics", 12 2010, http://hal.inria.fr/inria-00550651/en/.

[28] C. Raïssi, J. Pei, T. Kister. *Computing Closed Skycubes*, in "Proceedings of the VLDB Endowment (PVLDB)", 2010, vol. 3, n^{o} 1, p. 838–847.

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

[30] 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, p. 219–238.

[31] V. Venkatraman, V. Pérez-Nueno, L. Mavridis, D. Ritchie. *A Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods*, in "Journal of Chemical Information and Computer Sciences", November 2010 [*DOI :* 10.1021/CI100263P], http://hal.inria.fr/inria-00540762/en.

Articles in National Peer-Reviewed Journal

[32] R. Bendaoud, Y. Toussaint, A. Napoli. *L'Analyse Formelle de Concepts au service de la construction et l'enrichissement d'une ontologie*, in "Revue des Nouvelles Technologies de l'Information RNTI E-18", 2010, vol. Fouille de données complexes : avancées récentes, p. 133–163.

[33] E. NAUER, Y. TOUSSAINT. *Crechaindo : un système itératif et interactif de classification par treillis de concepts, pour la recherche d'information sur le web*, in "Document numérique", 2010, vol. 13, n$^{\text{o}}$ 1, p. 41–62 [*DOI :* 10.3166/DN.13.1.41-62], http://hal.inria.fr/inria-00526622/en.

International Peer-Reviewed Conference/Proceedings

[34] Z. ASSAGHIR, M. KAYTOUE, A. NAPOLI, H. PRADE. *Managing Information Fusion with Formal Concept Analysis*, in "7th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2010)", Perpignan, France, V. TORRA, Y. NARUKAWA, M. DAUMAS (editors), Lecture Notes in Computer Science 6408, Springer, Berlin, 2010, p. 104–115.

[35] A. BLANSCHÉ, J. COJAN, V. DUFOUR-LUSSIER, J. LIEBER, P. MOLLI, E. NAUER, H. SKAF-MOLLI, Y. TOUSSAINT. *TAAABLE 3: Adaptation of ingredient quantities and of textual preparations*, in "18h International Conference on Case-Based Reasoning - ICCBR 2010, "Computer Cooking Contest" Workshop Proceedings", Italy Alessandria, 2010, http://hal.inria.fr/inria-00526663/en.

[36] A. BLANSCHÉ, H. SKAF-MOLLI, P. MOLLI, A. NAPOLI. *Human-machine Collaboration for Enriching Semantic Wikis using Formal Concept Analysis*, in "Fifth Workshop on Semantic Wikis – Linking Data and People (SemWiki-2010)", C. LANGE, J. REUTELSHOEFER, S. SCHAFFERT, H. SKAF-MOLLI (editors), CEUR Workshop Proceedings Vol-632, 2010.

[37] J. CAO, P. KARRAS, C. RAÏSSI, K.-L. TAN. *rho-uncertainty: Inference-Proof Transaction Anonymization*, in "Proceedings of the VLDB Endowment (PVLDB)", 2010, vol. 3, n$^{\text{o}}$ 1, p. 1033–1044.

[38] J. COJAN, J. LIEBER. *An Algorithm for Adapting Cases Represented in an Expressive Description Logic*, in "18th International Conference on Case-Based Reasoning - ICCBR 2010", Italy Alessandria, I. BICHINDARITZ, S. MONTANI (editors), Lecture Notes in Artificial Intelligence, Springer Berlin, July 2010, vol. 6176, p. 51-65 [*DOI :* 10.1007/978-3-642-14274-1], http://www.springerlink.com/content/g314847630u162t7/?p=439be231431d441daea7fd617ad6bff3&pi=5, http://hal.inria.fr/inria-00506078/en.

[39] V. DUFOUR-LUSSIER, J. LIEBER, E. NAUER, Y. TOUSSAINT. *Text adaptation using formal concept analysis*, in "18th International Conference on Case-Based Reasoning - ICCBR 2010", Italy Alessandria, I. BICHINDARITZ, S. MONTANI (editors), Lecture Notes in Artificial Intelligence, Springer-Verlag, July 2010, vol. 6176, p. 96-110 [*DOI :* 10.1007/978-3-642-14274-1_9], http://hal.inria.fr/hal-00509030/en.

[40] P. EKLUND, J. VILLERD. *A Survey of Hybrid Representation of Concept Lattices in Conceptual Knowledge Processing*, in "I8th International Conference on Formal Concept Analysis - ICFCA 2010", L. KWUIDA, B. SERTKAYA (editors), Lecture Notes in Artificial Intelligence 5986, Springer, 2010, p. 296–311 [*DOI :* 10.1007/978-3-642-11928-6_21], http://www.springerlink.com/content/p180171m871h3752/, http://hal.inria.fr/inria-00503254/en.

[41] M. KAYTOUE, Z. ASSAGHIR, S. O. KUZNETSOV, A. NAPOLI. *Embedding Tolerance Relations in Formal Concept Analysis – An Application in Information Fusion*, in "Proceedings of CIKM 2010, 19th ACM Conference on Information and Knowledge Management", Toronto, Canada, 2010.

[42] M. KAYTOUE, Z. ASSAGHIR, N. MESSAI, A. NAPOLI. *Two Complementary Classication Methods for Designing a Concept Lattice from Interval Data*, in "Sixth International Symposium on Foundations of Information and Knowledge Systems (FoIKS)", Sofia, Bulgaria, S. LINK, H. PRADE (editors), Lecture Notes in Computer Science 5956, Springer, Berlin, 2010, p. 345–362.

[43] 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, p. 55–58.

[44] B. Lamiroy, Y. Guebbas. *Robust and Precise Circular Arc Detection*, in "8th IAPR International Workshop on Graphics RECognition - GREC 2009", France La Rochelle, J.-M. Ogier, W. Liu, J. Lladós (editors), Lecture Notes in Computer Science, Springer-Verlag, 2010, vol. 6020, p. 49-60 [*DOI :* 10.1007/978-3-642-13728-0_5], http://www.springerlink.com/index/K8VH03W872078412.pdf, http://hal.inria.fr/inria-00516712/en.

[45] B. Lamiroy, D. Lopresti. *A Platform for Storing, Visualizing, and Interpreting Collections of Noisy Documents*, in "Fourth Workshop on Analytics for Noisy Unstructured Text Data - AND'10", Canada Toronto, ACM International Conference Proceeding Series, ACM, October 2010 [*DOI :* 10.1145/1871840.1871844], http://hal.inria.fr/inria-00516678/en.

[46] E. G. Lazrak, M. Benoît, J.-F. Mari. *Time-Space Dependencies in Land-Use Successions at Agricultural Landscape Scales*, in "International Conference on Integrative Landscape Modelling", France Montpellier, Symposcience, February 2010, http://hal.inria.fr/inria-00482890/en.

[47] F. Le Ber, C. Lavigne, K. Adamczyk, F. Angevin, N. Colbach, J.-F. Mari, H. Monod. *Neutral modelling of agricultural landscapes by tessellation methods: the GenExP-LandSiTes software - Application to the simulation of gene flow*, in "International Conference on Integrative Landscape Modelling", France Montpellier, 2010, p. 1–9, http://hal.inria.fr/hal-00468772/en.

[48] X. Lu, W. H. Tok, C. Raïssi, S. Bressan. *A Simple, Yet Effective and Efficient, Sliding Window Sampling Algorithm*, in "Database Systems for Advanced Applications, 15th International Conference, DASFAA 2010", Tsukuba, Japan, H. Kitagawa, Y. Ishikawa, Q. Li, C. Watanabe (editors), Lecture Notes in Computer Science 5981, April 1-4 2010, p. 337–351.

[49] L. Mavridis, D. Ritchie. *3D-blast: 3D protein structure alignment, comparison, and classification using spherical polar Fourier correlations*, in "Pacific Symposium on Biocomputing 2010", United States Hawaii, World Scientific Publishing, January 2010, p. 281–292 [*DOI :* 10.1142/9789814295291_0030], http://hal.inria.fr/inria-00434263/en.

[50] L. Mavridis, V. Venkatraman, D. Ritchie, H. Morikawa, R. Andonov, A. Cornu, N. Malod-Dognin, J. Nicolas, M. Temerinac-Ott, M. Reisert, H. Burkhardt, A. Axenopoulos, P. Daras. *SHREC'10 Track: Protein Models*, in "Eurographics Workshop on 3D Object Retrieval - 3DOR 2010", Sweden Norrköping, 2010, http://hal.inria.fr/inria-00536680/en.

[51] N. Messai, M.-D. Devignes, A. Napoli, M. Smaïl-Tabbone. *Using Domain Knowledge to Guide Lattice-based Complex Data Exploration*, in "19th European Conference on Artificial Intelligence - ECAI 2010", Portugal Lisbon, H. Coelho, R. Studer, M. Wooldridge (editors), Frontiers in Artificial Intelligence and Applications, IOS press, 2010, vol. 215, p. 847–852, ISBN : 978-1-60750-605-8 [*DOI :* 10.3233/978-1-60750-606-5-847], http://hal.inria.fr/inria-00545545/en.

[52] A. Napoli. *Why and How Knowledge Discovery Can Be Useful for Solving Problems with CBR*, in "Case-Based Reasoning. Research and Development, 18th International Conference on Case-Based Reasoning, ICCBR 2010", Alessandria, Italy, I. Bichindaritz, S. Montani (editors), Lecture Notes in in Computer Science 6176, Springer, Berlin, July 19-22 2010, p. 12–19.

[53] D. Ritchie, V. Venkatraman, L. Mavridis. *Using Graphics Processors to Accelerate Protein Docking Calculations*, in "HealthGrid 2010", France Paris, 2010, http://hal.inria.fr/inria-00537989/en.

[54] M. Rouane-Hacene, M. Huchard, A. Napoli, P. Valtchev. *Using Formal Concept Analysis for Discovering Knowledge Patterns*, in "CLA'10: 7th International Conference on Concept Lattices and Their Applications", Spain Sevilla, M. Kryszkiewicz, S. Obiedkov (editors), CEUR Workshop Proceedings Vol-672, University of Sevilla, October 2010, p. 223–234, http://hal.inria.fr/lirmm-00531802/en.

[55] K. Santosh, C. Nattee, B. Lamiroy. *Spatial Similarity based Stroke Number and Order Free Clustering*, in "International Conference on Frontiers in Handwriting Recognition", India Kolkata, 2010, http://hal.inria.fr/inria-00516726/en.

[56] K. Santosh, L. Wendling, B. Lamiroy. *Using Spatial Relations for Graphical Symbol Description*, in "20th International Conference on Pattern Recognition - ICPR 2010", Turkey Istanbul, IEEE, 2010, p. 2041 - 2044, ISSN: 1051-4651<br /> Print ISBN: 978-1-4244-7542-1 [*DOI :* 10.1109/ICPR.2010.503], http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5595915, http://hal.inria.fr/inria-00516725/en.

[57] 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, p. 16–27.

[58] J. Villerd, Y. Toussaint, A. L.-L. Louët. *Adverse Drug Reaction Mining in Pharmacovigilance Data Using Formal Concept Analysis*, in "Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010", Barcelona, Spain, J. L. Balcázar, F. Bonchi, A. Gionis, M. Sebag (editors), Lecture Notes in Computer Science 6323, Springer, 2010, p. 386–401.

National Peer-Reviewed Conference/Proceedings

[59] Z. Assaghir, M. Kaytoue, A. Napoli, H. Prade. *Organisation de la fusion d'information avec l'analyse formelle de concepts*, in "Rencontres francophones sur la Logique Floue et ses Applications (LFA)", Cépaduès-Éditions, 2010, p. 133–140.

[60] Z. Assaghir, M. Kaytoue, A. Napoli, H. Prade. *Organisation de la fusion d'information avec l'Analyse Formelle de Concepts*, in "Journées d'Intelligence Artificielle Fondamentale – IAF 2010", Strasbourg, France, L. Cholvy, S. Konieczny (editors), 2010.

[61] J. Cojan, J. Lieber. *Adapter des cas en utilisant un opérateur de révision ou des règles*, in "Journée Intelligence Artificielle Fondamentale", France Strasbourg, L. Cholvy, S. Konieczny (editors), Laurence Cholvy, Sébastien Konieczny, June 2010, http://gdri3iaf.info.univ-angers.fr/spip.php?article121, http://hal.inria.fr/inria-00512529/en.

[62] J. Cojan, J. Lieber. *Un algorithme d'adaptation avec des cas exprimés dans la logique de descriptions ALC*, in "18ème Atelier « Raisonnement à Partir de Cas » RàPC 2010", France Strasbourg, F. L. Ber, J. Renaud (editors), June 2010, p. 37-48, http://hal.inria.fr/inria-00506094/en.

[63] V. Dufour-Lussier, J. Lieber, E. Nauer, Y. Toussaint. *Améliorer la remémoration par enrichissement de l'ontologie du domaine*, in "Atelier RàPC", France Strasbourg, 2010, http://hal.inria.fr/inria-00526639/en.

[64] V. Dufour-Lussier, J. Lieber, E. Nauer, Y. Toussaint. *Enrichir une ontologie pour améliorer la recherche d'information approximative*, in "Atelier "Evolution d'ontologies" lors de la conférence Ingénierie des Connaissances (IC) 2010", France Nîmes, 2010, http://hal.inria.fr/inria-00526641/en.

[65] B. Fuchs, M. Huchard, A. Napoli. *Une étude sur la mise en forme de patrons de conception pour les ontologies avec l'analyse formelle de concepts*, in "LMO 2010 : Langages et Modèles à Objets", Pau, France, E. Cariou, J.-C. Royer (editors), Université de Pau et des Pays de l'Adour, March 2010, p. 83–98, http://hal.inria.fr/lirmm-00534516/en.

[66] C. Grac, A. Braud, F. Le Ber, M. Trémolières. *Un système d'information pour le suivi et l'évaluation de la qualité des cours d'eau*, in "Actes du 3ème atelier " Systèmes d'Information et de Décision pour l'Environnement" - SIDE 2010 - Congrès Inforsid", France Marseille, 2010, p. 12–21, http://hal.inria.fr/hal-00498545/en.

[67] M. Kaytoue, Z. Assaghir, N. Messai, A. Napoli. *Classification de données numériques par treillis de concepts basée sur une similarité symbolique/numérique*, in "Comptes-rendus des journées de la SFC", Saint Denis de La Réunion, 2010, p. 99–102.

[68] M. Kaytoue, Z. Assaghir, N. Messai, A. Napoli. *Complémentarité de deux méthodes de classification pour la construction de treillis de concepts à partir de données numériques*, in "Actes du 17ème Congrès Francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA 2010)", Caen, France, M.-O. Cordier, J.-M. Jolion (editors), AFRIF–AFIA, 2010, p. 399–406.

[69] S. Lardon, F. Le Ber. *Cas socio-spatiaux pour le diagnostic prospectif de territoires*, in "18ème Atelier Raisonnement à Partir de Cas - RàPC 2010", France Strasbourg, 2010, p. 83-90, http://hal.inria.fr/hal-00497574/en.

[70] F. Le Ber, J. Lieber, A. Napoli. *Représentation temporelle qualitative de recettes de cuisine*, in "RTE 2010 - atelier associé à la conférence RFIA 2010", France, 2010, p. 25–28, http://hal.inria.fr/hal-00459322/en.

[71] J.-F. Mari, E. G. Lazrak, M. Benoît. *Fouille de paysages agricoles: analyse des voisinages des successions d'occupation du sol*, in "Colloque RTE (Raisonnement sur le Temps et l'Espace) en marge de RFIA 2010", France Caen, M. Bouzid, F. L. Ber, G. Ligozat, O. Papini (editors), RFIA 2010, January 2010, http://hal.inria.fr/inria-00482811/en.

[72] N. Messai, M.-D. Devignes, A. Napoli, M. Smaïl-Tabbone. *Connaissances de domaine et treillis de concepts pour l'exploration progressive de données complexes*, in "21es Journées francophones d'Ingénierie des Connaissances – IC 2010", France Nîmes, M. Crampes, S. Desprès (editors), Ecole des Mines d'Alès, 2010, p. 233–244, http://hal.inria.fr/hal-00488034/en.

[73] J. Villerd, Y. Toussaint, A. L.-L. Louët. *Complémentarité des méthodes numériques et symboliques en pharmacovigilance*, in "21èmes Journées Francophones d'Ingénierie des Connaissances", France Nîmes, S. Desprès (editor), Ecole des Mines d'Alès, 2010, p. 221-232, http://hal.inria.fr/inria-00506606/en.

Scientific Books (or Scientific Book chapters)

[74] F. L. Ber, J. Renaud (editors). *18ème Atelier « Raisonnement à Partir de Cas » RàPC 2010*, 2010, http://hal.inria.fr/hal-00497210/en.

[75] K. SANTOSH, L. WENDLING, B. LAMIROY. *Unified Pairwise Spatial Relations: An Application to Graphical Symbol Retrieval*, in "Graphics Recognition. Achievements, Challenges, and Evolution", J.-M. OGIER, W. LIU, J. LLADÓS (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2010, vol. 6020, p. 163-174 [*DOI :* 10.1007/978-3-642-13728-0_15], http://hal.inria.fr/inria-00516724/en.

Books or Proceedings Editing

[76] M. BOUZID, F. LE BER, G. LIGOZAT, O. PAPINI (editors). *Actes du 5ème atelier Représentation et raisonnement sur le temps et l'espace (RTE 2010)*, 2010, http://hal.inria.fr/hal-00542329/en.

[77] F. LE BER, T. LIBOUREL (editors). *Informations géographiques – Connaissances et enjeux environnementaux*, Hermès - Lavoisier, 2010, Numéro spécial de "Revue internationale de géomatique", ISSN 1260-5875, vol. 20, n° 2, http://hal.inria.fr/hal-00516979/en.

Research Reports

[78] M. KAYTOUE, S. O. KUZNETSOV, Z. ASSAGHIR, A. NAPOLI. *Embedding Tolerance Relations in Concept Lattices - An application in Information Fusion*, INRIA, August 2010, n^{o} RR-7353, http://hal.inria.fr/inria-00508462/en.

[79] M. KAYTOUE, S. O. KUZNETSOV, A. NAPOLI. *Pattern Mining in Numerical Data: Extracting Closed Patterns and their Generators*, INRIA, October 2010, n^{o} RR-7416, http://hal.inria.fr/inria-00526662/en.

Other Publications

[80] B. DUCATEL, M. KAYTOUE, F. MARCUOLA, A. NAPOLI, L. SZATHMARY. *CORON : Plate-forme d'Extraction de Connaissances dans les Bases de Données*, in "Démonstrations, 17ème Congrès Francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA 2010)", Caen, France, M.-O. CORDIER, J.-M. JOLION (editors), AFRIF–AFIA, Caen, France, 2010.

[81] T. MEILENDER, N. JAY, J. LIEBER, F. PALOMARES. *Les moteurs de wikis sémantiques : un état de l'art*, 2010, http://hal.inria.fr/hal-00542813/en.

[82] N. SCHALLER, E. G. LAZRAK, P. MARTIN, J.-F. MARI, C. AUBRY, M. BENOÎT. *Modelling regional Land Use: articulating the farm and the landscape levels by combining farmers' decision rules and landscape stochastic regularities*, August 2010, Poster session, European Society of Agronomy, Agropolis2010, Montpellier.

References in notes

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

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

[85] P. HITZLER, M. KRÖTSCH, S. RUDOLPH (editors). *Foundations of Semantic Web Technologies*, CRC Press, Boca raton (FL), 2009.

[86] S. STAAB, R. STUDER (editors). *Handbook on Ontologies (Second Edition)*, Springer, Berlin, 2009.

[87] F. BADRA, R. BENDAOUD, R. BENTEBITEL, P.-A. CHAMPIN, J. COJAN, A. CORDIER, S. DESPRÈS, S. JEAN-DAUBIAS, J. LIEBER, T. MEILENDER, A. MILLE, E. NAUER, A. NAPOLI, Y. TOUSSAINT. *Taaable: Text Mining, Ontology Engineering, and Hierarchical Classification for Textual Case-Based Cooking*, in "ECCBR 2008, The 9th European Conference on Case-Based Reasoning, Trier, Germany, September 1-4, 2008, Workshop Proceedings", 2008, p. 219-228.

[88] F. BADRA, J. COJAN, A. CORDIER, J. LIEBER, T. MEILENDER, A. MILLE, P. MOLLI, E. NAUER, A. NAPOLI, H. SKAF-MOLLI, Y. TOUSSAINT. *Knowledge acquisition and discovery for the textual case-based cooking system WIKITAAABLE*, in "8th International Conference on Case-Based Reasoning - ICCBR 2009, Workshop Proceedings", Seattle États-Unis d'Amérique, S. J. DELANY (editor), 07 2009, p. 249–258, http://hal.inria.fr/inria-00411508/en/.

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

[90] 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, p. 311–323.

[91] 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, p. 156–171.

[92] 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, p. 203–216.

[93] J. BERNAUER, J. AZÉ, J. JANIN, A. POUPON. *A new protein-protein scoring function based on interface residue properties*, in "Bioinformatics", 2007, vol. 23, p. 555–562.

[94] C. CARPINETO, G. ROMANO. *Concept Data Analysis: Theory and Applications*, John Wiley & Sons, Chichester, UK, 2004.

[95] C. CARPINETO, G. ROMANO. *Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO.*, in "Journal of Universal Computer Science", 2004, vol. 10, n^{o} 8, p. 985–1013.

[96] A. CARRIERI, V. PÉREZ-NUENO, A. FANO, C. PISTONE, D. RITCHIE, J. TEIXIDÓ. *Biological Profiling of Anti-HIV Agents and Insight into CCR5 Antagonist Binding Using in silico Techniques*, in "ChemMedChem", 2009, vol. 4, p. 1153–1163, http://dx.doi.org/10.1002/cmdc.200900101.

[97] 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, p. 305–339.

[98] J. COJAN, J. LIEBER. *Belief Merging-based Case Combination*, in "8th International Conference on Case-Based Reasoning - ICCBR 2009 Case-Based Reasoning Research and Development", Seattle États-Unis

d'Amérique, D. C. WILSON, L. MCGINTY (editors), Lecture Notes in Computer Science, Springer Berlin, 07 2009, vol. 5650, p. 105–119, http://hal.inria.fr/inria-00421724/en/.

[99] C. ENG. *Développement de méthodes de fouille de données fondées sur les modèles de Markov cachés du second ordre pour l'identification d'hétérogénéités dans les génomes bactériens*, Université Henri Poincaré Nancy 1, 2010.

[100] 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, p. 129–142.

[101] B. GANTER, R. WILLE. *Formal Concept Analysis*, Springer, Berlin, 1999.

[102] Y. GARTEN. *Text mining the scientific literature to identify pharmacogenomic interactions*, Stanford University, USA, Dec 2010.

[103] N. JAY, F. KOHLER. *Comment évaluer la qualité des données de bases régionales ou nationales des RSA en vue d'une utilisation épidémiologique à partir de l'analyse formelle de concept et des treillis de Galois*, in "Actes des XXII journées EMOIS", 2009.

[104] S. K.C., B. LAMIROY, J.-P. ROPERS. *Inductive Logic Programming for Symbol Recognition*, in "Tenth International Conference on Document Analysis and Recognition - ICDAR'2009", Barcelona Spain, IEEE, 07 2009, p. 1330 - 1334, INRIA-INPL, http://hal.inria.fr/inria-00430927/en/.

[105] B. LAMIROY, J.-P. ROPERS. *Assessing Inductive Logic Programming Classification Quality by Image Synthesis*, in "Eighth IAPR International Workshop on Graphics Recognition - IAPR-GREC 2009", La Rochelle France, J.-M. OGIER, L. WENYIN, J. LLADOS (editors), University of La Rochelle, 07 2009, p. 344-352, http://hal.inria.fr/inria-00439456/en/.

[106] M. F. LENSINK, S. J. WODAK. *Docking and scoring protein interactions: CAPRI 2009*, in "Proteins", 2010, vol. 78, p. 3073–2084.

[107] 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, p. 261–274.

[108] 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, p. 22–41.

[109] J.-F. MARI, F. L. BER. *Temporal and Spatial Data Mining with Second-Order Hidden Models*, in "Soft Computing", 2006, vol. 10, n^{o} 5, p. 406–414.

[110] J.-F. MARI, J.-P. HATON, A. KRIOUILE. *Automatic Word Recognition Based on Second-Order Hidden Markov Models*, in "IEEE Transactions on Speech and Audio Processing", 1997, vol. 5, p. 22 – 25.

[111] R. MOSCA, C. PONS, J. FERNANDEZ-RECIO, P. ALOY. *Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments*, in "PLoS Computational Biology", 2009, vol. 5, n^{o} 8, e1000490 [*DOI :* 10.1371/JOURNAL.PCBI.1000490].

[112] 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, p. 913–933.

[113] C. A. ORENGO, A. D. MICHINE, S. JONES, D. T. JONES, M. B. SWINDELLS, J. M. THORNTON. *CATH - A Hierarchic Classification of Protein Domain Structures*, in "Structure", 1997, vol. 5, n^{o} 8, p. 1093–1108.

[114] 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^{o} 4, p. 810-823 [*DOI :* 10.1021/CI800468Q], http://hal.inria.fr/inria-00434261/en.

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

[116] V. PÉREZ-NUENO, D. RITCHIE, O. RABAL, R. PASCUAL, J. BORRELL, J. TEIXIDÓ. *Comparison of ligand-based and receptor-based virtual screening of HIV entry inhibitors for the CXCR4 and CCR5 receptors using 3D ligand shape matching and ligand-receptor docking*, in "Journal of Chemical Information and Modeling", 2008, vol. 48, n^{o} 3, p. 509–533.

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

[118] 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, p. 51–65.

[119] G. R. SMITH, M. J. E. STERNBERG. *Prediction of protein-protein interactions by docking methods*, in "Current Opinion in Structural Biology", 2002, vol. 12, n^{o} 1, p. 28-35.

[120] A. STEIN, R. B. RUSSELL, P. ALOY. *3did: interacting protein domains of known three-dimensional structure*, in "Nucleic Acids Res.", 2005, vol. 33, p. D413–D417.

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

[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, p. 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, p. 393–404.

[124] C. WINTER, A. HENSCHEL, W. KIM, M. SCHROEDER. *SCOPPI: a structural classification of protein-protein interfaces*, in "Nucleic Acids Research", 2006, vol. 34, p. D310–D314.

[125] T. YOSHIKAWA, K. TSUKAMOTO, Y. HOURAI, K. FUKUI. *Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins*, in "Journal of Chemical Information and Modeling", 2009, vol. 49, p. 693–703.

[126] M. D'AQUIN, F. BADRA, S. LAFROGNE, J. LIEBER, A. NAPOLI, L. SZATHMARY. *Case Base Mining for Adaptation Knowledge Acquisition*, in "Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07)", M. M. VELOSO (editor), Morgan Kaufmann, 2007, p. 750–755.