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

Table of contents

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

1. Team

Research Scientist

Amedeo Napoli [Team leader, Researcher (DR CNRS), HdR]
Marie-Dominique Devignes [Researcher (CR CNRS), HdR]
Bernard Maigret [Researcher (DR CNRS), HdR]
Chedy Raissi [Researcher (CR INRIA, since November 1st)]
Dave Ritchie [Chaire Excellence ANR]
Yannick Toussaint [Researcher (CR INRIA)]

Faculty Member

Nicolas Jay [Associate Professor (Faculté de Médecine, UHP Nancy 1)]
Florence Le Ber [Professor (ENGEES Strasbourg), HdR]
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)]
Frédéric Pennerath [PhD Student (Thesis defended in July 2009) and lecturer (Supélec Metz)]
Malika Smaïl-Tabbone [Associate Professor (MdC Université Henri Poincaré Nancy 1)]

Technical Staff

Thomas Meilender [Engineer]
Alexandre Blansché [Engineer (since November 1st)]
Florent Marcuola [Engineer]
Birama NDiayé [Engineer]

PhD Student

Zainab Assaghir [PhD Student (INRA Grant)]
Yasmine Assess [PhD Student (INCa Grant)]
Fadi Badra [PhD Student (ATER, Thesis defended in October 2009)]
Sid-Ahmed Benabderrahmane [PhD Student (INCa Grant)]
Emmanuel Bresso [Master Student (February – September 2009, PhD Student (Harmonic Pharma Grant)]
Rokia Bendaoud [PhD Student (ATER, until August 31th)]
Matthieu Chavent [PhD Student (CNRS-Région Grant, until June 30th)]
Julien Cojan [PhD Student (AMX Grant)]
Léo Ghemtio [PhD Student (ANR Contract)]
Anisha Ghoorah [PhD Student (ANR Contract, from April 2009)]
Mehdi Kaytoue [PhD Student (MERT Grant)]
Nizar Messaï [PhD Student (ATER, until August 31th)]
Thomas Meilender [Engineer and PhD Student since October 2009)]

Post-Doctoral Fellow

Vincent Leroux [Post-Doctoral fellow (INCa Grant)]
Lazaros Mavridis [Post-Doctoral fellow (ANR Grant, from March 2009)]
Vishwesh Venkatraman [Post-Doctoral fellow (ANR Grant, from July 2009)]
Jean Villerd [Post-Doctoral fellow (ANR Project)]

Administrative Assistant

Emmanuelle Deschamps [Secretary]

Other

Zaina Hassanzadeh [Master Student (from February until September 2009)]
Ahmed Zeeshan [Master Student (from February until June 2009)]

2. Overall Objectives

2.1. Introduction

Knowledge discovery in databases –hereafter KDD– consists in processing a large volume of data in order to extract 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: the "orpailleur" denotes in French 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 an *ontology* 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, knowledge units that are extracted have still a life after the interpretation step: they must be represented in an adequate knowledge representation formalism for being integrated within an ontology and reused for problem-solving needs. In this way, the results of the knowledge discovery process may be reused for extending and updating existing ontologies. The KDDK process shows that knowledge representation and knowledge discovery are two complementary tasks: *no effective knowledge discovery without domain knowledge!*

2.2. Highlights

This is the second year that the Taaable system is currently developed for participating in a challenge, namely the "Computer Cooking Contest". The Taaable system is available on line at http://taaable.fr. This system has been designed with the collaboration of the SILEX team (LIRIS Lyon) and the RCLN team (LIPN Paris 13). The Taaable system won the second prize in the first "Computer Cooking Contest" [69] at the European Conference on Case-Based Reasoning in Trier (Germany, September 2008). In addition, the system has also won the the second prize in the second "Computer Cooking Contest" (International Conference on Case-Based Reasoning, July 2009, Seattle, USA) [35]. The design of the Taaable system involved a large part of the Orpailleur team, and needed joint efforts and combination of many skills and capabilities, such as knowledge representation, ontology engineering, classification, case-based reasoning, text-mining, information retrieval.

3. Scientific Foundations

3.1. From KDD to KDDK

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 *data mining methods* that are either symbolic or numerical. The methods that are used in the Orpailleur team are the following:

- Symbolic methods are based on lattice-based classification (or concept lattice design or formal concept analysis [78]), frequent itemsets search, and association rule extraction [85].
- Numerical methods based on second-order Hidden Markov Models (HMM2, designed for pattern recognition [84]). Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data.

Then, the principle summarizing KDDK can be read as follows [82]: going "from complex data units to complex knowledge units guided by domain knowledge" (KDDK) or "knowledge with/for knowledge". Two original aspects can be underlined: (i) the fact that the KDD process is guided by domain knowledge, and (ii) the fact that the extracted units are embedded within a knowledge representation formalism to be reused in a knowledge-based system for problem solving purposes.

In the research work of the Orpailleur team, the various instantiations of the KDDK process are all based on the idea of *classification*. Classification is a polymorphic process involved in various tasks, e.g. modeling, mining, representing, and reasoning. Accordingly, a knowledge-based system may be designed, fed up by the KDDK process, and used for problem-solving in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine, with a special mention for semantic web activities involving text mining, content-based document mining, and intelligent information retrieval [67], [68].

3.2. Methods for Knowledge Discovery guided by Domain Knowledge

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. Lattice-based classification 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 concepts organized within a hierarchy (i.e. a partial ordering) [70]. Lattice-based classification is used for building concept lattices, also called Galois lattices, and is the basic operation underlying the so-called *formal concept analysis* or FCA [78].

The search for frequent itemsets and association rule extraction are well-known symbolic data mining methods, related to lattice-based classification. These 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 (major [10]) [94].

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.

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 [83]). 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

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 [66], [74] [62]. 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 manipulating textual documents by their content.

The studies on text mining carried out in the Orpailleur team hold on real-world texts in application domains such as astronomy, biology and medicine, using mainly symbolic data mining methods [13]. This is in contrast with text analysis approaches dealing with specific language phenomena. The language in texts is considered as a way for presenting and accessing information, and not as an object to be studied for its own. Accordingly, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

3.4. Elements on Knowledge Systems and Semantic Web

Knowledge representation is a process for representing knowledge within a knowledge representation formalism, giving knowledge units a syntax and a semantics. The **semantic web** is a framework for building knowledge-based systems for manipulating documents on the web by their contents, i.e. taking into account 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. Nowadays, the web is becoming "semantic" in the sense that people search for information with the help of machines, that are in charge of asking questions, searching for answers, classifying and interpreting the answers. The web becomes a space for exchange of information between machines, allowing an "intelligent access" and "management" of information. 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 setting up a 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 associated with the semantic web is the OWL language, based on description logics (or DL [65]). 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". Concept classification is used for inserting a new concept at the right location in the concept hierarchy, searching for its most specific subsumers and its most general subsumees.

Individual classification is used for recognizing the concepts an individual may be an instance of. Furthermore, classification-based reasoning may be extended into 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. When there is enough interest, the target problem and its solution may be memorized in the case base to be reused. In the context of a concept hierarchy, retrieval and adaptation may be both based on classification and adaptation-guided retrieval [76].

4. Application Domains

4.1. Life Sciences

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Matthieu Chavent, Marie-Dominique Devignes, Léo Gemthio, Mehdi Kaytoue, Vincent Leroux, Nizar Messai, Bernard Maigret, Amedeo Napoli, Malika Smaïl-Tabbone, Yannick Toussaint.

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

One of the major application domains currently investigated by the Orpailleur team is related to life sciences, with a particular emphasis on biology (bioinformatics), medicine, and chemistry. The understanding of biological systems provides complex problems for computer scientists, and, when problems are solved, solutions bring new ideas not only for biologists but also for computer scientists. Moreover, the team includes biologists, chemists, and a physician, making Orpailleur a very original INRIA team.

Knowledge discovery is gaining more and more interest and importance in life sciences for mining either homogeneous databases such as protein sequences or structures, 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 may involve 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: Fadi Badra, Julien Cojan, Jean Lieber, Thomas Meilender, Amedeo Napoli.

The Kasimir research project holds on decision support and knowledge management for the treatment of cancer [81]. 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 Hermès (an association for the sharing of resources in informatics for medicine). 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. In addition, protocol adaptations are studied from the ergonomics and computer science viewpoints. These adaptations can be used to propose *evolutions* of the protocol based on a confrontation with actual cases. The idea is then to make suggestions for protocol evolutions based on frequently performed adaptations.

Adaptation plays a central role in knowledge-intensive CBR, where a target problem is solved by adapting the solution of a source case. The adaptation process is based on adaptation knowledge that –for the main part– is domain-dependent, and thus needs to be acquired for a new application of CBR. While the acquisition of ontologies is one important issue that is widely explored in the semantic web community, the acquisition of decision and adaptation knowledge has not been so deeply explored, though this kind of knowledge can be useful in numerous situations. Accordingly, this is the goal of *adaptation knowledge acquisition* (AKA) to mine a case base, to extract adaptation knowledge units, and to make these units operational. The AKA process is aimed at feeding a knowledge server embedded in the Kasimir semantic portal, that includes an OWL-based formalisms for representing medical ontologies, decision protocols (the case base), and adaptation knowledge [96]. Web services associated to the CBR process are developed, and several protocols are implemented.

5. Software

5.1. KDD Systems

5.1.1. The Coron Platform

Keywords: *association rule extraction*, *data mining*, *frequent closed itemsets*, *frequent generators*, *frequent itemsets*, *rare itemsets*.

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

The Coron platform [93] 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 extract different kinds of itemsets, e.g. frequent itemsets, frequent closed itemsets, frequent generators, etc. The algorithms are APriori, APriori-Close, Close, Pascal, Eclat, Charm, and, as well, original algorithms such as Pascal+, ZART, Carpathia, Eclat-Z, and Charm-MFI. AssRuleX (Association Rule eXtractor) generates different sets of association rules (from itemsets), e.g. minimal non-redundant association rules, generic basis, informative basis, etc. The Coron-base component contains also algorithms for extracting rare itemsets and rare association rules, e.g. APriori-rare, MRG-EXP, ARIMA, and BTB.

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 module RuleMiner facilitates the interpretation and the filtering of the extracted rules. The association rules can be filtered by (i) attribute, (ii) support, and/or (iii) confidence. It is also possible to color the most important attributes in the list of rules, for finding the most interesting rules from a given viewpoint.

The Coron toolkit is developed entirely in Java, is operational, and has already been used within several research projects, e.g. for mining the Stanislas cohort, or in the CabamakA project (which is part of the Kasimir system, see §4.2). An extension of the system, named BioCoron, is aimed at taking into account gene expression [80].

5.1.2. The CarottAge system

Keywords: *Hidden Markov Models*, *stochastic process*.

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

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

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

In applications, the systems 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).

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.

5.1.3. GenExp-LandSiTes: KDD and simulation

Keywords: *Hidden Markov Models*, *Simulation*.

Participants: 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 [57] (major [7]). The last version of GenExp allows an interaction with R subroutines and has received a GPL License.

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.

5.1.4. KDD systems in Biology

Participants: Marie-Dominique Devignes [contact person], Nizar Messai, Malika Smaïl-Tabbone, Marie-Dominique Devignes [contact person], Birama Ndiaye, Malika Smaïl-Tabbone, Marie-Dominique Devignes [contact person], Bernard Maigret, Malika Smaïl-Tabbone.

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

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 searchable 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 formal concept analysis allows a user to browse and discover databases through the BioRegistry. A publication on this work has been accepted in the International Journal of Metadata, Semantics and Ontology. The BioRegistry repository is available at http://bioregistry.loria.fr.

MOdel-driven Data Integration for Mining (MODIM).

A position of "Ingénieur Jeune Diplomé INRIA" has been 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. MODIM was inspired by a previous work on an Approach for Candidate Gene Retrieval (ACGR) (major [11]).

Graphical interface for the Virtual Screening platform (Virtual Screening Manager for the computing grid: VSM-G).

The graphical interface for the virtual screening platform VSM-G is currently in used and declared as an INRIA APP at the beginning of 2009.

5.2. Knowledge-Based Systems and Semantic Web

5.2.1. CreChainDo

Keywords: *association rule extraction*, *frequent itemset search*, *information retrieval*, *knowledge discovery from databases*, *navigation*, *semantic web*, *text mining*.

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

The "CreChainDo" second system makes use of FCA for information retrieval on the web. Many recent systems use FCA for improving the access to documents on the web. Among them, the Credo system [72], [73], uses a concept lattice to reorganize the list of documents returned by a search engine as an answer to a given query. In Credo, a lattice is built according to the title and the snippet of each documents returned by Google. Navigating into the lattice hierarchy guides the access to the web documents.

In this way, a lattice contains concepts that are relevant and some others that are not relevant for a given information retrieval task. Extending the Credo approach, we introduce lattices into an interactive and iterative system, called CRECHAINDO (major [8]). The CRECHAINDO system uses FCA for reorganizing the list of documents returned by Google according a lattice. The lattice, presented as a tree-hierarchy, helps the user to explore the search results in a structured and synthetic way. The CRECHAINDO system offers to the user a way of expressing a negative or positive agreement with some concept of the lattice, in agreement with the objective of information retrieval. These user choices are converted into extension or reduction operations on the lattice, in order to make the lattice evolve and to better fit his/her needs.

5.2.2. The Kasimir System for Decision Knowledge Management

Keywords: *case-based reasoning*, *classification-based reasoning*, *decision knowledge management*, *edition and maintenance of knowledge*, *semantic portal*.

Participants: Fadi Badra, Julien Cojan, 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. Actually, two versions of Kasimir are currently used. A first version is based on an *ad hoc* object-based representation formalism. A second version is developed within a semantic portal, based on OWL and extensions of OWL, implying the development of the two user interfaces, namely EdHibou and NavHibou [97]. The instance editor EdHibou is used for querying the protocols represented within the Kasimir system. The browser NavHibou is developed for navigating in the class hierarchies built by a reasoner based on OWL. Moreover, since the Kasimir inference engine is based on subsumption, a study on the integration of an extended inference engine taking into account inferences based on CBR, and on an integration within the semantic web, is under study.

The software CabamakA for case base mining for adaptation knowledge acquisition is a module of the Kasimir system [12]. This system performs case base mining for adaptation knowledge acquisition and provides information units to be used for building adaptation rules [96]. Actually, the mining process in CabamakA is implemented thanks to a frequent close itemset extraction module of the Coron platform (see §5.1.1). The adaptation knowledge acquisition process is not fully automated: an analyst guides CabamakA, following the principles of knowledge discovery, i.e. the analyst filters and interprets the results of the mining process, to be rewritten into adaptation rules.

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

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

Participants: Fadi Badra, Julien Cojan, Jean Lieber [contact person], Thomas Meilender, 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.

The Taaable system is available on line at http://taaable.fr. This system has been designed with the collaboration of the SILEX team (LIRIS Lyon) and the RCLN team (LIPN Paris 13). In addition, Taaable won the second price in the first "Computer Cooking Contest" [69] (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) [35].

6. New Results

6.1. The Mining of Complex Data

Participants: Zainab Assaghir, Rokia Bendaoud, Nicolas Jay, Mehdi Kaytoue, Florence Le Ber, Amedeo Napoli, Frédéric Pennerath, Yannick Toussaint.

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, n-ary relations, graphs, texts, etc.

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. graph-based or relational data. Among them, Relational Concept Analysis (RCA) is a process for analyzing objects described both by binary and relational attributes [91]. 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 [71].

Another extension of FCA is based on so-called Pattern Structures (PS) [77], 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). A research work is currently under development on the relations existing between classification methods based on FCA with explicit similarity measure (Formal Concept Analysis driven by Similarity or FCAS [14], [33], [60] and Pattern Structure classification. The parallel study of SCA and PS helps to understand how these two methods are interrelated and how they can be applied to complex data for building concept lattices.

PC classification and FCAS have been applied in the field of decision support in agronomy. There exists a set of agro-ecological indicators aimed at helping farmers to improve their agricultural practices. Actually, an indicator estimates the impact of cultivation practices on the "agrosystem" [79]. The modeling and the assessment of environmental risk generally require a large number of parameters whose measure is imprecise. Thus, it is important to study how imprecision is propagated in the various steps of decision support, and, as well, which are the different types of imprecision that are combined in the computation of the value of an indicator [58]. This is not really straightforward, but the computation of an indicator value, decision support based on indicators (that cab be seen as a special decision tree), and pattern structure classification, are all linked and can be studied in the same classification framework.

Still in the context of agronomy, a series of 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 design of concept lattices for mining and understanding complex hydrobiological data, requiring specific algorithms [48], [37], [49], [38]. These studies are of general interest as they try to push forward the computational capabilities of standard FCA algorithms by considering complex data with multiple nested modalities.

For completing the work on FCA, there is still on-going work on frequent itemset search for improving standard algorithms, but also for being able to build lattices from very large data. In this case, closed itemsets are searched for first, then generators, i.e. minimal itemsets in their equivalent classes, and finally the association of each equivalent classes between each other, giving in fact the concept ordering in the underlying concept lattice (major [10]).

6.1.2. KDDK in Medico-Economical Databases

Since 30 years, many patient classification systems (PCS) have been developed. Theses 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 essentially used for funding purposes, it holds potentially very useful knowledge for other public health domains such as epidemiology or health care planning. Our main objective is to extract knowledge units from this database in order to explore "Patient Care Trajectories". Our approach aims at assisting domain experts with automated classification tools to define or to detect particular groups of patients having similar health condition, treatments or journeys through the healthcare system. To achieve these tasks, we propose a methodology based on Formal Concept Analysis (FCA). From a theoretical point of view, our research focuses on the ability of FCA to deal with large amounts of data. We especially study means of reducing complexity of large concept lattices for easy visualization and selection of the most interesting results. Our methods have been applied for data quality assessment of the PMSI in epidemiology [53] and diagnostic strategies comparison [27].

Another way of research consists in data driven ontology building. The idea is to reuse knowledge discovered during the FCA step for providing an ontology of PCT that will perform reasoning tasks on patient profiles. Such an ontology could, for example, help to qualify a chronic disease made of a succession of pathological states.

6.1.3. 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.

A preliminary research work has been carried on in the Orpailleur team, based on frequent levelwise itemset search and association rule extraction, and applied to standard chemical reaction databases. Given the results of this work, a subsequent research has been carried out involving this time a graph-mining process used for extracting knowledge from chemical reaction databases, directly from the molecular structures and the reactions themselves.

This research work is currently under development, in collaboration with chemists and in accordance with needs of chemical industry. This year, once more, a number of substantial results have been obtained and presented in some high-level conferences (major [9]) [30], [15].

6.2. KDDK and Text Mining

Participants: Rokia Bendoaud, Amedeo Napoli, Emmanuel Nauer, Yannick Toussaint, Jean Villerd.

The objective of text mining is to extract useful and reusable knowledge units from large collections of texts. An objective of the team is to make available extracted knowledge units for allowing a "machine-based" manipulation of texts.

6.2.1. Knowledge discovery from heterogeneous textual resources

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 for ontology engineering from heterogeneous ontological resources. A methodology and a system, called "Pactole", have been designed and have been applied in various contexts, namely in astronomy and in biology [13]. The "Pactole" methodology extends previous research works based on FCA and aimed at building ontologies from ontological resources using formal concept analysis and relational concept analysis.

The "Pactole" methodology is based on the identification in texts of objects, and on the extraction of object properties and of relations between objects. Object identification is possible thanks to a list of names (for example the celestial object "HR2725" or the bacteria "Echerichia Coli") or a set of patterns ("NGC xxxx" where "xxxx" is a number). Properties and relations between objects are extracted from the texts using syntactic parsers (e.g. Stanford parser) and information extraction tools (e.g. Gate). Properties are expressed in texts with adjectives or verbs while relations are usually expressed through lexical patterns.

Then, binary tables "Objects $\times$ Attributes" are built and the associated concept lattices can be computed. In addition, a transformation function may convert the lattice into a concept hierarchy expressed in a simple description logic formalism (FLE). The RCA process has been used to take into account relations between objects and to create relation between concepts of the ontology. Moreover, an interactive process based on FCA and RCA including the analyst into the KDD loop when building an ontology has been studied.

Meanwhile, beside ontology engineering, a survey on the use of association rules for text mining, mainly for classifying extracted association rules from texts, has been published, giving a conclusion to this research aspect in the team [62].

6.2.2. KDDK in Pharmacovigilance

Participants: Yannick Toussaint, Jean Villerd.

Pharmacovigilance (PV) holds on 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. A *safety signal* appears when a single drug consumption is the cause of an (unexpected) ADR. A *syndrome* appears when a single drug consumption is the cause of several (unexpected) ADRs. A *drug interaction* appears when the consumption of several drugs is the cause of an (unexpected) ADR. A *protocol* appears when the consumption of several drugs is the cause of several (unexpected) ADRs.

The ANR Project Vigitermes was running its second year in 2009. The primary goal of this project is 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.

We developed first an approach based on association rules [44]. However, trying to establish a better formulation of expert needs led us to propose a new method for identifying candidates for pharmacological associations to be investigated in clinical trials. A clinical trial allows the observation of a drug activity on a given population. The identification method relies on Formal Concept Analysis. The lattice resulting from FCA is used as a "search space" for searching patterns in itemsets associated to concepts in the lattice. The subsumption relation between concepts in the lattice is used to relate signals, interactions, and protocols (as introduced above). In addition, this identification method uses several statistical components for numerically filtering significant associations. The method has been implemented within a prototype system and validated through an experiment on a data base from the "Georges Pompidou Hospital".

6.3. Current Research and New Perspectives in Life Sciences

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Emmanuel Bresso, Matthieu Chavent, Marie-Dominique Devignes, Léo Gemthio, Anisha Ghoorah, Mehdi Kaytoue, Florence Le Ber, Vincent Leroux, Bernard Maigret, Jean-François Mari, Lazaros Mavridis, Nizar Messai, Amedeo Napoli, Dave Ritchie, Vishwesh Venkatraman, Malika Smaïl-Tabbone.

6.3.1. KDDK in Life Sciences

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.

6.3.1.1. Virtual Screening

Virtual screening (VS) techniques are nowadays widely recognized as interesting techniques as part of early drug discovery strategies, since when successful they provide an excellent cost-to-efficiency ratio. In a high-throughput screening context (millions of candidates), VS techniques are still under-exploited. In particular, the popular molecular docking programs are either too slow or considered as not reliable enough compared to more expensive experimental protocols. One way to overcome such limitations involves coupling multiple techniques in a funnel-like filtering process. Several filtering strategies can be set up in this context such as in VSM-G software. VSM-G uses as large-scale first filtering step a crude geometrical docking algorithm based on spherical harmonics. We have studied a knowledge-oriented approach that could complement this algorithm in reducing the number of false positives. The rationale of this approach is that extracting patterns from data relative to known active compounds can be used to filter out inactive compounds from chemical libraries. This approach was tested on the Liver X receptor (LXR), on the Apelin receptor, and on the C-Met receptor, which are all targets of interest [18], [19], [20], [23], [24] (major [1]).

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

Virtual screening has become an essential step in the early drug discovery process. It consists in using computational techniques for selecting compounds from chemical libraries in order to identify drug-like molecules acting on a biological target of therapeutic interest. We consider virtual screening as a particular form of KDD process. The knowledge units to be discovered concern the way a compound can be considered as a consistent ligand for a given target. The data from which knowledge has to be discovered derive from diverse sources such as chemical, structural, and biological data related to ligands and their cognate targets. More precisely, an objective is to extract "filters" from chemical libraries and protein-ligand interactions. Three basic steps of a KDD process have been implemented. Firstly, a model-driven data integration step is applied to appropriate heterogeneous data found in public databases. This facilitates subsequent extraction of various datasets to be mined. In particular and for specific ligand descriptors, it allows transforming a multiple-instance

problem into a single-instance one. In a second step, mining algorithms were applied to datasets and finally the most accurate knowledge units are assessed as new virtual screening filters. The experimental results obtained with a set of ligands of the hormone receptor LXR have been published in [41].

6.3.1.3. Knowledge Discovery from Transcriptomic Data

This work concerns the interpretation of transcriptomic data from colorectal cancer samples and is the subject of an ongoing PhD thesis funded by the INCa (Institut National du Cancer) in collaboration with Olivier Poch (IGBMC, Strasbourg). DNA microarray technologies allow to monitor the expression of several thousands of genes in different situations. The expression levels measured for each gene in a set of situations define a gene expression profile. Usually a functional analysis is then applied to genes with similar expression profiles. We have proposed a new approach based on a priori modeling of Differential Expression Profiles (DEP) considering the relations between the situations. Fuzzy logic is used for assigning genes to DEPs. Results with data on colorectal cancer show that this modeling of DEP lead to relating biological functions with defined transcriptional behavior [47].

Further functional analysis of DEPs requires a flexible gene-gene similarity measure that takes into account at best domain knowledge mostly represented here by the annotation vocabulary known as Gene Ontology (GO). Various semantic similarity measures exist [87] that consider both semantic relationships between annotation terms and their information content. However none of them includes yet the quality of gene annotations which is reported as evidence codes in the public databases. We are currently testing a new similarity measure, that is defined in a vectorial framework inspired from information retrieval and considers both semantic relationships of terms, their information content, and quality metadata.

6.3.1.4. Relational data mining applied to 3D protein patches for characterizing and predicting phosphorylation sites

An ongoing study is in concern with the prediction of phosphorylation sites through the design of models exploiting information on the 3D structure of proteins and methods of logical relational data mining based on Inductive Logic Programming (ILP) [75]. Indeed, relational data mining appears as a relevant way to extract knowledge units from 3D structures and the prediction of phosphorylation sites constitutes a well-documented case-study. During the nineties, several ILP success stories were reported on biological problems concerning the prediction of protein 3D structure starting from the protein primary sequence. An idea is to use here the same ILP methods to further predict biological phenomena. We are motivated by testing the ability of ILP techniques to provide explicit insights about the considered biological problem. Current results reveal interesting features of what constitutes a phosphorylation site in terms of predicates describing the 3D patch surrounding this site.

6.3.1.5. Using FCA for analyzing biological data

FCA is the basic classification method used in two research topics: (i) classification of biological DBs on the Web, (ii) analysis and classification of Gene Expression Data (GED). In the first track, the BioRegistry project aims at organizing metadata about biological DBs in order to ease classification and retrieval tasks. Metadata do not have the same importance and the same structure. Thus, two main extensions of FCA have been designed. The first allows the introduction of dependencies between attributes, e.g. attribute hierarchies, which are considered for concept lattice construction. The second is aimed at handling many-valued contexts: the so-called SimBA for "Similarity-Based Complex Data Analysis System" algorithm builds a many-valued concept lattice using similarity between attribute values. The results of these two extensions of FCA are substantial and are detailed in [14], and, as well, have given birth to new research perspective on similarity and pattern structures as explained in § 6.1.1.

Another research work involving FCA holds on the analysis of gene expression data (GED) for discovering groups of co-expressed genes [54]. Microarray biotechnology is able to measure the expression of a gene (related to its activity) in a given biological situation. A gene expression profile (GEP) is considered as a numerical m-dimensional vector, describing the behavior of the gene. A gene expression data (GED) is a collection of n gene expression profiles and is represented as an $n \times m$ numerical table. FCA is applied for analyzing and interpreting the data, knowing that genes having a similar expression profile may participate

in a same biological process. Accordingly, formal concepts in the resulting concept lattice are representing sets of genes presenting similar variations of expression in biological situations. Substantial results have been obtained applying FCA to a real dataset related to the fungus "Laccaria Bicolor" for studying interaction between fungus and poplars (a very important tree in the industry of wood).

6.3.1.6. Mining Biological Data with HMMs

In this particular research direction for KDDK, we have designed a new data mining method based on stochastic analysis (Hidden Markov Model or HMM) and combinatorial methods for discovering new transcriptional factors in bacterial genome sequences (major [5]). Sigma factor binding sites (SFBSs) were described as patterns corresponding to DNA motifs of bacterial promoters. High-order HMM are used in which the hidden process is a second-order HMM chain and applied to the genome of bacterium Streptomyces coelicolor and Bacillus subtilis. Short DNA sequences were extracted by HMM and clustered with a hierarchical classification algorithm. Some selected motif consensuses were combined with over-represented motifs found by a word enumeration algorithm. This original and new mining methodology applied to several genomes was able to retrieve SFBSs and to suggest new potential transcriptional factor binding sites.

In another investigation field, namely agricultural landscapes, methods for identifying and describing meaningful landscape patterns play an important role to understand the interaction between landscape organization and ecological processes. We have proposed an innovative stochastic modeling method of agricultural landscape organization where the temporal regularities in land-use are first identified through recognized Land-Use Successions before locating these successions in landscapes [25], [25]. These time-space regularities within landscapes are extracted using a data mining method based on HMM. We applied this method to the Niort Plain (West of France). Implications and perspectives of such an approach, which links together the temporal and the spatial dimensions of the agricultural organization, have been investigated by assessing the relationship between the agricultural landscape patterns and ecological issues.

6.3.2. Structural Systems Biology

6.3.2.1. High Performance Algorithms for Structural Systems Biology (HPASSB)

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.

6.3.2.2. 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. 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-protein systems. We will present this work as a poster entitled "Fast FFT Protein-Protein Docking on Graphics Processors" at the 4th CAPRI Evaluation Meeting in Barcelona in December 2009.

6.3.2.3. 3D-Blast: A New Approach for Protein Structure Alignment and Clustering

We have recently developed a new sequence-independent protein structre alignment approach, which we call 3D-Blast, based on the spherical polar Fourier (SPF) correlation approach used in the *Hex* protein docking software [90]. The utility of this approach has been demonstrated by clustering subsets of the CATH protein structure classification database [86] 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 the 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 [28].

6.3.2.4. 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 [95] and 3DID [92] protein interaction databases to help make some very good predictions to two or 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 will be presented as a poster at the 4th CAPRI Evaluation Meeting in Barcelona in December 2009.

6.4. Around the Kasimir and Taaable research projects

Participants: Fadi Badra, Julien Cojan, Jean Lieber, Thomas Meilender, Amedeo Napoli.

6.4.1. CabamakA and Adaptation Knowledge Acquisition

The research about adaptation within the Kasimir research project is described in [81]. Adaptation in Kasimir, as well as in many other CBR systems, requires knowledge. The adaptation knowledge acquisition (AKA) is a research work, that can take two directions: AKA from experts (manual) and semi-automatic AKA (using KDD). AKA from experts consists in analyzing adaptations performed by experts. Interviews of experts confronted to decision problems requiring adaptation are analyzed and modeled as adaptation patterns.

Semi-automatic AKA is based on the mining of protocol rules "`situation`$\longrightarrow$`decision`". Knowing how the decisions change when the situations change from one rule to another rule provides a specific adaptation rule. By generalizing these specific rules, general adaptation rules may be obtained. This generalization process can be implemented through a frequent closed itemset extraction algorithm. The system called CabamakA realizes the mining of protocol rules for adaptation rule acquisition [96]. The KDD process in CabamakA is based on the Coron platform (see §5.1.1). Moreover, an analyst guides the AKA process in CabamakA, using filters to drive the mining process, and interpreting the extracted pieces of information in adaptation rules [12].

AKA from experts and semi-automatic AKA are not completely satisfying: The "AKA from experts" provides generic adaptation patterns that are intelligible, but cannot be directly operational. The "semi-automatic AKA" provides adaptation rules that can be directly implemented, but that are difficult to understand (and thus, to validate). A possible research work would be to combine these two kinds of AKA for producing operational and intelligible adaptation knowledge units.

6.4.2. New Directions in the Taaable Project

The Taaable project has been originally created as a challenger of the Computer Cooking Contest (CCC, http://www.wi2.uni-trier.de/ccc09). 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) 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. It has been presented as a demo in [51]. The second version (2009) was the World vice champion of the contest. A third version for the 2010's contest is under conception.

The partners of the 2009's Taaable project are members of Orpailleur, of the SCORE Team at LORIA, and of the SILEX team of the LIRIS (Lyon). 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, ontology engineering, semantic wikis, text-mining, etc.

A general description of the 2009's Taaable system, also called WikiTaaable, can be found in [35]. The most important original features of this version are:

- The use of a semantic wiki architecture for the collaborative edition of the Taaable knowledge [40], [50]: formalized recipes, a cooking ontology, adaptation and retrieval knowledge;
- Opportunistic adaptation knowledge acquisition and extraction [12], [46] (major [2]): an inadequate result of the system triggers a process of knowledge acquisition and extraction which results in new pieces of adaptation knowledge. The extraction process is based on the CabamakA system which uses a data-mining algorithm of the Coron platform.

The current case-based reasoning inference engine of Taaable and what should be its future are described in [56]. Another ongoing work is about the application of minimal change theory to adaptation in case-based reasoning and its future application to Taaable [59]. (major [4]). The main contribution in this area in 2009 is the application of this approach to a formalism including quantitative values and how it can be reduced, under some assumptions, to linear programming. Finally, a research about temporal reasoning for case-based reasoning with an application to the adaptation of recipes is under study [55].

7. Other Grants and Activities

7.1. International projects and collaborations

7.1.1. The AmSud Project: Semantic-based support for Collaborative Design Activity

Participants: Alexandre Blansché, Amedeo Napoli [contact person], Yannick Toussaint.

The main goal of this cooperation project is to define a methodological and software support for integrating semantic technologies in a collaborative computer-supported design activity. The project intends to demonstrate that semantic web technologies are a suitable and efficient option for improving collaborative computer-assisted design process. Thus, semantic requirements for design situations have to be identified, and, methodological support and software solution for management have to be developed.

There is a need for studying how collaborative design activities can be guided by domain knowledge. Here, domain ontologies have to help designers in assembling design components, in searching adequate components, detecting conflicts, searching related documents, finding people with the adequate skills, etc. The design work can be done by several people distributed over time, space, and organizations. Wikis are very good examples of such distributed activities. Moreover, a special attention is given to semantic wikis: semantic activities for wiki design and wiki management for semantic activities.

This project involves researchers from LORIA (Orpailleur and SCORE teams), LIFIA at Universita de La Plata (Argentina), Laboratório Intermidia at Universidade de Sao Paulo in Sao Carlos (Brazil), and Departamento de Informática de Universidad Técnica Federico Santa María (Santiago do Chile, Valparaiso). Two publications are associated to the project [34], [61].

7.1.2. International Collaborations in Biology and Chemistry

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Emmanuel Bresso, Matthieu Chavent, Marie-Dominique Devignes, Léo Gemthio, Anisha Ghoorah, Vincent Leroux, Bernard Maigret, Lazaros Mavridis, Nizar Messai, Dave Ritchie, Vishwesh Venkatraman, Malika Smaïl-Tabbone.

7.1.2.1. 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" and ran from September 2008 until September 2009. A joint publication is currently in preparation.

7.1.2.2. 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 Sarria (IQS) in Barcelona, Dave Ritchie published a review of the state of *in silico* protein structure modeling and *virtual drug screening* techniques regarding the development of CCR5 entry-blockers (major [3]). As there now exist several hundred CCR5 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 [89]. In September, Dave Ritchie gave a presentation at the international Modeling-09 conference in Erlangen on recent collaborative work with Violeta Pérez-Nueno of the IQS on clustering and classifying novel CCR5 HIV entry inhibitors using spherical harmonic shape representations [88]. 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 CCR4 [31].

7.1.2.3. 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 (Target 34: our predictions were amongst the best for this target according to the ligand RMSD criteria ; Target 40: we identified key residues before they were revealed by the organisers thanks to litterature and analogues interactions). It was presented as a poster at the 4th CAPRI Evaluation Meeting in Barcelona in December 2009.

7.1.3. Other international collaborations

Participants: Mehdi Kaytoue, Amedeo Napoli.

Two close international research collaborations have to be still mentioned: (i) one with Petko Valtchev at Université du Québec à Montréal (UQAM) and Marianne Huchard at LIRMM Montpellier, (ii) the other with Sergei Kusnetsov at Higher School of Economics in Moscow (HSE). The two collaborations are based on visits in the different laboratories and in the writing of common papers. The research topic in (i) holds on the design of algorithms for itemset search and association rule extraction, and the design of large concept lattices and relational concept lattices (major [10]). The research topic in (ii) holds on extension of FCA algorithms for taking into account complex data (major [6]).

7.2. National initiatives

7.2.1. ANR Nutrivigène

Participants: Mehdi Kaytoue, Florent Marcuola, Amedeo Napoli.

Nutrigenomics is an emerging topic interested in elaborating dietary recommendations and new food products. The Nutrivigène project aims at studying "homocysteine" that is closely correlated with age and associated with vascular, cognitive, and neurological dysfunctions. Homocystéine is an intermediate product of the carbon metabolism and is also related to the status of folate and vitamin B12. 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 (people recruited in a rural region of Sicily). The originality of this study consists in identifying the nutrigenomic mechanisms related to homocysteine and its nutritional and genetic determinants that may alter the epigenetic of the expression of genes involved in the vascular, cognitive and neurological functional deterioration of aged people. The evaluation of the association between the methylation of gene candidates (MS-PCR) and the vascular and cognitive function is carried using KDDK methods.

The Nutrivigène project involves the following partners: INSERM U724 (Nancy Hospital) in association with IRCCS of Troina (Italy), INRA Alimentation Humaine (Clermont-Ferrand Theix), UMR CNRS 2738 (Marseille), ERI 11-INSERM Nancy, LSGA (INPL Nancy), LORIA (Orpailleur Team), and finally Nestlé-Waters (Vittel).

7.2.2. ANR Vigitermes: Mining for signal in Pharmacovigilance

Participants: Yannick Toussaint, Jean Villerd.

Pharmacovigilance covers research activities related to detection, analysis, and prevention of unexpected adverse drug reactions (ADR). In France, health-care professionals have to declare serious or unexpected ADRs, while spontaneous reports can be collected in two different ways. The pharmacovigilance units of pharmaceutical laboratories receive spontaneous reports on ADRs that concern the drugs they commercialize. The regional "Pharmacovigilance Centers" collect spontaneous reports on ADRs for all drugs commercialized in France. These reports are registered in the pharmacovigilance national database, AFSSaPS for "Agence Française de Sécurité Sanitaire des Produits de Santé". At the international level, the WHO for "World Health Organization" Program, which was established in 1968, consists of a network of National Centers, WHO Headquarters, Geneva, and the WHO Collaborating Center for International Drug Monitoring, the Uppsala Monitoring Center (UMC, Uppsala, Sweden). Individual case reports of suspected adverse drug reactions are collected and stored in a common database, presently containing over 3.7 million case reports, in several languages.

The general objective of the Vigitermes project consists in supporting the work of Pharmacovigilance experts –in pharmaceutical industry and regulatory agencies– in two ways: firstly, in their interaction with available resources, e.g. Pharmacovigilance database, product catalogs, medical literature; secondly, in improving signal detection in pharmacovigilance. This work is based on KDDK methods including knowledge management in pharmacovigilance.

Vigitermes involves 10 partners: DSPIM (Department of Public Health and Medical Informatics of Saint-Etienne), SPIM (INSERM U872, équipe 20), Modélisation Conceptuelle des Connaissances Biomédicales (EA 3888), LIM&BIO (Laboratoire d'Informatique Médicale & Bio-Informatique, EA3969), CRPV of the European Georges Pompidou Hospital (HEGP), TEMIS, Mondeca, INRIA-Orpailleur, and Inalco. The World Health Organisation Uppsala Monitoring Centre for Drug Safety (WHO-UMC) is an associated partner.

7.3. Local initiatives

The link between the Regional Administration and LORIA is materialized within the so-called "Contrat Plan État Région" (CPER). This contract is named "Modélisations, informations et systèmes numériques" (MISN) and runs from 2007 until 2013. In the global research contract, there are two research projects in which the Orpailleur team is involved.

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

This project is coordinated by M.-D. Devignes (http://bioinfo.loria.fr). The general goal of this research project is to study how domain knowledge can be taken into account for improving modeling of biomolecules and their interactions, and how, in turn, this helps the modeling of biological systems. Six projects involving collaborations with biology or chemistry laboratories are currently under development.

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

  Research operations are currently under development on knowledge management and decision support in the large involving the Kasimir and the Taaable systems.

8. Dissemination

8.1. Scientific Animation

- The members of the Orpailleur team are involved, as members or as head persons, in a number of national research groups.
- The members of the Orpailleur team are involved in the organization of conferences, as members of conference program committees, as members of editorial boards, and finally in the organization of journal special issues.

8.2. Teaching

- The members of the Orpailleur team are involved in teaching at all levels of teaching in the universities of Nancy, especially "Université Henri Poincaré Nancy-1" and "Université de Nancy-2"; actually, it must be noticed that 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.

9. Bibliography

Major publications by the team in recent years

[1] Y. ASSES, V. LEROUX, S. TAIRI-KELLOU, R. DONO, F. MAINA, B. MAIGRET. *Analysis of c-Met Kinase Domain Complexes: A New Specific Catalytic Site Receptor Model for Defining Binding Modes of ATP-Competitive Ligands*, in "Chemical Biology & Drug Design", vol. 74, 2009, p. 560-570, http://hal.inria.fr/inria-00435159/en/.

[2] F. BADRA, A. CORDIER, J. LIEBER. *Opportunistic Adaptation Knowledge Discovery*, in "8th International Conference on Case-Based Reasoning, ICCBR 2009 ICCBR, Seattle États-Unis d'Amérique", L. MCGINTY, D. C. WILSON (editors), Lecture Notes in Computer Science, vol. 5650, Springer, 07 2009, p. 60-74, http://hal.inria.fr/inria-00437693/en/.

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

[4] 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, vol. 5650, Springer Berlin, 07 2009, p. 105-119, http://hal.inria.fr/inria-00421724/en/, The original publication is available at www.springerlink.com.

[5] C. ENG, C. ASTHANA, B. AIGLE, S. HERGALANT, J.-F. MARI, P. LEBLOND. *A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods*, in "Journal of Computational Biology", vol. 16, 09 2009, p. 1211-1225, http://hal.inria.fr/inria-00419969/en/, J.: Computer Applications/J.3: LIFE AND MEDICAL SCIENCES/J.3.0: Biology and genetics Région Lorraine, CPER-MBI.

[6] M. KAYTOUE-UBERALL, S. DUPLESSIS, S. KUZNETSOV, A. NAPOLI. *Two FCA-Based Methods for Mining Gene Expression Data*, in "Formal Concept Analysis – Proceedings of the 7th International Conference – ICFCA 2009, Darmstadt, Germany", S. FERRÉ, S. RUDOLF (editors), Lecture Notes in Artificial Intelligence 5548, Springer, Berlin, 2009, p. 251–266.

[7] F. LE BER, C. LAVIGNE, K. ADAMCZYK, F. ANGEVIN, N. COLBACH, J.-F. MARI, H. MONOD. *Neutral modelling of agricultural landscapes by tessellation methods – Application for gene flow simulation*, in "Ecological Modelling", vol. 220, 2009, p. 3536–3545.

[8] E. NAUER, Y. TOUSSAINT. *CreChainDo: an iterative and interactive Web information retrieval system based on lattices*, in "International Journal of General Systems", vol. 38, 2009, p. 363-378, http://hal.archives-ouvertes.fr/inria-00336845/en/.

[9] F. PENNERATH, A. NAPOLI. *The Model of Most Informative Patterns and its Application to Knowledge Extraction from Graph Databases*, in "European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2009, Bled, Slovenia", W. BUNTIME, M. GROBELNIK, J. SHAWE-TAYLOR, D. MLADENIC (editors), Lecture Notes in Artificial Intelligence 5782, Springer, Berlin, 2009, p. 205-220.

[10] 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.

[11] S. YILMAZ, P. JONVEAUX, C. BICEP, L. PIERRON, M. SMAIL-TABBONE, M.-D. DEVIGNES. *Gene-Disease Relationship Discovery based on Model-driven Data Integration and Database View Definition*, in "Bioinformatics", vol. 25, 2009, p. 230–236.

Year Publications

Doctoral Dissertations and Habilitation Theses

[12] F. BADRA. *Extraction de connaissances d'adaptation en raisonnement à partir de cas*, Université Henri Poincaré - Nancy I, 11 2009, http://tel.archives-ouvertes.fr/tel-00438140/en/, Ph. D. Thesis.

[13] R. BENDAOUD. *Analyses formelle et relationnelle de concepts pour la construction d'ontologies de domaines à partir de ressources textuelles hétérogènes*, Université Henri Poincaré - Nancy I, 2009, Ph. D. Thesis.

[14] N. MESSAI. *Extraction de connaissances et Web sémantique. Application à la recherche et l'intégration de ressources génomiques sur le Web*, University Henri Poincaré – Nancy 1, France, March 2009, Thèse d'Université en informatique.

[15] F. PENNERATH. *Méthodes d'extraction de connaissances à partir de données modélisables par des graphes. Application à des problèmes de synthèse organique*, Université Henri Poincaré Nancy 1, 2009, Thèse d'informatique.

Articles in International Peer-Reviewed Journal

[16] Y. ASSES, V. LEROUX, S. TAIRI-KELLOU, R. DONO, F. MAINA, B. MAIGRET. *Analysis of c-Met Kinase Domain Complexes: A New Specific Catalytic Site Receptor Model for Defining Binding Modes of ATP-Competitive Ligands*, in "Chemical Biology & Drug Design", vol. 74, 2009, p. 560-570, http://hal.inria.fr/inria-00435159/en/.

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

[18] C. CLAPERON, I. BANEGAS-FONT, X. ITURRIOZ, R. ROZENFELD, B. MAIGRET, C. LLORENS-CORTES. *Identification of threonine 348 as a residue involved in aminopeptidase A substrate specificity*, in "The Journal of Biological Chemistry", vol. 284, n^{o} 16, 2009, p. 10618-26, http://www.hal.inserm.fr/inserm-00418798/en/.

[19] E. DE OLIVEIRA, C. HUMEAU, L. CHEBIL, E. MAIA, F. DEHEZ, B. MAIGRET, M. GHOUL, J. ENGASSER. *A molecular modelling study to rationalize the regioselectivity in acylation of flavonoïd glycosides catalysed by Candida antartica lipase B*, in "Journal of Molecular Catalysis B: Enzymatic", vol. 59, 2009, p. 96-105, http://hal.inria.fr/inria-00435075/en/.

[20] N. DÉLIOT, M. CHAVENT, C. NOURRY, P. LÉCINE, C. ARNAUD, A. HERMANT, B. MAIGRET, J. BORG. *Biochemical studies and Molecular Dynamics Simulations of Smad3-Erbin interaction identify a non-classical Erbin PDZ binding*, in "Biochemical and Biophysical Research Communications / Biochemistry and Biophysics Research Communications", vol. 378, 2009, p. 360-365, http://hal.inria.fr/inria-00339128/en/.

[21] C. ENG, C. ASTHANA, B. AIGLE, S. HERGALANT, J.-F. MARI, P. LEBLOND. *A new data mining approach for the detection of bacterial promoters combining stochastic and combinatorial methods*, in "Journal of Computational Biology", vol. 16, 09 2009, p. 1211-1225, http://hal.inria.fr/inria-00419969/en/, J.: Computer Applications/J.3: LIFE AND MEDICAL SCIENCES/J.3.0: Biology and genetics Région Lorraine, CPER-MBI.

[22] S. FERRARESSO, H. KUHL, M. MILAN, D. W. RITCHIE, C. J. SECOMBES, R. REINHARDT, L. BARGELLONI. *Identification and characterisation of a novel immune-type receptor (NITR) gene cluster in the European sea bass, Dicentrarchus labrax, reveals recurrent gene expansion and diversification by positive selection*, in "Immunogenetics", 2009, http://dx.doi.org/10.1007/s00251-009-0398-3, To Appear.

[23] N. FLOQUET, P. DURAND, B. MAIGRET, B. BADET, M.-A. BADET-DENISOT, D. PERAHIA. *Collective motions in glucosamine-6-phosphate synthase: influence of ligand binding and role in ammonia channelling and opening of the fructose-6-phosphate binding site*, in "Journal of Molecular Biology", vol. 385, n^{o} 2, 2009, p. 653-64, http://hal.archives-ouvertes.fr/hal-00365799/en/.

[24] A. KHALFA, W. TREPTOW, B. MAIGRET, M. TAREK. *Self assembly of peptides near or within membranes using coarse grained MD simulations*, in "Chemical Physics", vol. 358, 2009, p. 161-170, http://hal.inria.fr/inria-00435164/en/.

[25] E. G. LAZRAK, J.-F. MARI, B. MARC. *Landscape regularity modelling for environmental challenges in agriculture*, in "Landscape Ecology", 09 2009, http://hal.inria.fr/inria-00419952/en/.

[26] F. LE BER, C. LAVIGNE, K. ADAMCZYK, F. ANGEVIN, N. COLBACH, J.-F. MARI, H. MONOD. *Neutral modelling of agricultural landscapes by tessellation methods – Application for gene flow simulation*, in "Ecological Modelling", vol. 220, 2009, p. 3536–3545.

[27] I. LEMONNIER, C. BAUMANN, N. JAY, K. ALZAHOURI, P. ARVEUX, D. JOLLY, C. LEJEUNE, M. VELTEN, F. VITRY, M.-C. WORONOFF-LEMSI, F. GUILLEMIN. *Does the availability of positron emission tomography modify diagnostic strategies for solitary pulmonary nodules? An observational study in France*, in "BMC Cancer", vol. 9, 2009, 139, http://dx.doi.org/10.1186/1471-2407-9-139.

[28] L. MAVRIDIS, D. W. RITCHIE. *3D-Blast: Protein Protein Structure Alignment, Comparison, and Classification Using Spherical Polar Fourier Correlations*, in "Pacific Symposium on Biocomputing", vol. 2010, 2009, p. 281–292, http://dx.doi.org/10.1142/9789814295291_0030.

[29] E. NAUER, Y. TOUSSAINT. *CreChainDo: an iterative and interactive Web information retrieval system based on lattices*, in "International Journal of General Systems", vol. 38, 2009, p. 363-378, http://hal.archives-ouvertes.fr/inria-00336845/en/.

[30] F. PENNERATH, A. NAPOLI. *La famille des motifs les plus informatifs. Application à l'extraction de graphes en chimie organique*, in "Revue I3", vol. 8, n^{o} 2, 2009.

[31] V. PÉREZ-NUENO, S. PETTERSSON, D. W. RITCHIE, J. I. BORRELL, J. TEIXIDÓ. *Discovery of Novel HIV Entry Inhibitors for the CXCR4 Receptor by Prospective Virtual Screening*, in "Journal of Chemical Information and Modeling", vol. 49, 2009, p. 810–823, http://dx.doi.org/10.1021/ci800468q.

[32] S. YILMAZ, P. JONVEAUX, C. BICEP, L. PIERRON, M. SMAIL-TABBONE, M.-D. DEVIGNES. *Gene-Disease Relationship Discovery based on Model-driven Data Integration and Database View Definition*, in "Bioinformatics", vol. 25, 2009, p. 230–236.

International Peer-Reviewed Conference/Proceedings

[33] Z. ASSAGHIR, M. KAYTOUE, N. MESSAI, A. NAPOLI. *On the mining of numerical data with Formal Concept Analysis and similarity*, in "Book of Short Papers – Meeting of CLADAG, Catania (Italy)", S. INGRASSIA (editor), Edizione Scientifiche Italiane, Napoli, 2009, –, Also presented at Journées SFC, Grenoble septembre 2009.

[34] H. ASTUDILLO, G. CANALS, A. DIAZ, A. NAPOLI, M. PIMENTEL. *From data to knowledge through collaboration: bridging Wikis and Knowledge Systems*, in "COLIBRI 2009. Colloque d'informatique Brésil/Inria, Bento Gonçalves, Brésil", Universidade Federal Rio Grande do Sul, Porto Alegre, 2009, p. 211–217, http://hal.inria.fr/inria-00434397/en/.

[35] 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/.

[36] F. BADRA, A. CORDIER, J. LIEBER. *Opportunistic Adaptation Knowledge Discovery*, in "8th International Conference on Case-Based Reasoning, ICCBR 2009 ICCBR, Seattle États-Unis d'Amérique", L. MCGINTY, D. C. WILSON (editors), Lecture Notes in Computer Science, vol. 5650, Springer, 07 2009, p. 60-74, http://hal.inria.fr/inria-00437693/en/.

[37] A. BERTAUX, F. LE BER, A. BRAUD, M. TRÉMOLIÈRES. *Identifying ecological traits: a concrete FCA-based approach*, in "7th International Conference on Formal Concept Analysis, ICFCA 2009, Darmstadt", S. FERRÉ, S. RUDOLPH (editors), LNAI 5548, Springer-Verlag, 2009, p. 224–236.

[38] A. BRAUD, C. GRAC, S. PRISTAVU, E. DOR, F. LE BER. *Une démarche fondée sur les treillis de Galois pour l'aide à la qualification de l'état des milieux aquatiques*, in "Actes du 2ème Atelier « Systèmes d'Information et de Décision pour l'Environnement » – SIDE 2009, Toulouse, 26 mai 2009", 2009, p. 94–105.

[39] 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, vol. 5650, Springer Berlin, 07 2009, p. 105–119, http://hal.inria.fr/inria-00421724/en/, The original publication is available at www.springerlink.com.

[40] A. CORDIER, J. LIEBER, P. MOLLI, E. NAUER, H. SKAF-MOLLI, Y. TOUSSAINT. *WIKITAAABLE: A semantic wiki as a blackboard for a textual case-based reasoning system*, in "SemWiki 2009 - 4rd Semantic Wiki Workshop at the 6th European Semantic Web Conference - ESWC 2009, Heraklion Grèce", 05 2009, http://hal.inria.fr/inria-00432353/en/.

[41] L. GHEMTIO, M. SMAIL-TABBONE, M.-D. DEVIGNES, M. SOUCHET, B. MAIGRET. *A KDD Approach for Designing Filtering Strategies to Improve Virtual Screening*, in "KDIR - International Conference on Knowledge Discovery and Information Retrieval, Madeira Portugal", A. FRED (editor), INSTIC, 2009, http://hal.inria.fr/inria-00433475/en/.

[42] M. KAYTOUE-UBERALL, S. DUPLESSIS, S. KUZNETSOV, A. NAPOLI. *Two FCA-Based Methods for Mining Gene Expression Data*, in "Formal Concept Analysis – Proceedings of the 7th International Conference – ICFCA 2009, Darmstadt, Germany", S. FERRÉ, S. RUDOLF (editors), Lecture Notes in Artificial Intelligence 5548, Springer, Berlin, 2009, p. 251–266.

[43] F. PENNERATH, A. NAPOLI. *The Model of Most Informative Patterns and its Application to Knowledge Extraction from Graph Databases*, in "European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD 2009, Bled, Slovenia", W. BUNTIME, M. GROBELNIK, J. SHAWE-TAYLOR, D. MLADENIC (editors), Lecture Notes in Artificial Intelligence 5782, Springer, Berlin, 2009, p. 205-220.

[44] M. ROUANE HACENE, Y. TOUSSAINT, P. VALTCHEV. *Mining Safety Signals in Spontaneous Report Database using Concept Analysis*, in "12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona Italie", A. A.-H. CARLO COMBI (editor), Springer Berlin / Heidelberg, 07 2009, p. 285-294, http://hal.archives-ouvertes.fr/inria-00437224/en/.

[45] 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.

National Peer-Reviewed Conference/Proceedings

[46] F. BADRA, A. CORDIER, J. LIEBER. *Découverte opportuniste de connaissances d'adaptation*, in "17ème atelier sur le raisonnement à partir de cas - RàPC-09, Paris France", B. FUCHS, A. NAPOLI (editors), 06 2009, p. 23–34, http://hal.inria.fr/inria-00411509/en/.

[47] S. BENABDERRAHMANE, M.-D. DEVIGNES, M. SMAÏL-TABBONE, O. POCH, A. NAPOLI, N. NGUYEN N.-H, W. RAFFELSBERGER. *Analyse de données transcriptomiques: Modélisation floue de profils d'expression différentielle et analyse fonctionnelle*, in "Actes du XXVIIième congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 2009, Toulouse France", IRIT-Toulouse, 2009, p. 413-428, http://hal.inria.fr/inria-00394530/en/.

[48] A. BERTAUX, F. LE BER, A. BRAUD. *Correspondances de Galois pour la manipulation de contextes flous multi-valués*, in "Extraction et Gestion de Connaissances, EGC'2009, Strasbourg", J.-G. GANASCIA, P. GANÇARSKI (editors), RNTI E-15, Cépaduès, 2009, p. 193–198.

[49] A. BERTAUX, F. LE BER, P. LI, M. TRÉMOLIÈRES. *Combiner treillis de Galois et analyse factorielle multiple pour l'analyse de traits biologiques*, in "Actes des XVIèmes Rencontres de la Société Francophone de Classification, Grenoble", G. D'AUBIGNY (editor), septembre 2009, p. 117–120.

[50] A. CORDIER, J. LIEBER, P. MOLLI, E. NAUER, H. SKAF-MOLLI, Y. TOUSSAINT. *WikiTaaable, un wiki sémantique utilisé comme un tableau noir dans un système de raisonnement à partir de cas textuel*, in "17ème atelier de Raisonnement à Partir de Cas, Paris France", B. FUCHS, A. NAPOLI (editors), 06 2009, http://hal.archives-ouvertes.fr/inria-00437399/en/.

[51] A. CORDIER, J. LIEBER, E. NAUER, Y. TOUSSAINT. *Taaable : système de recherche et de création, par adaptation, de recettes de cuisine*, in "Extraction et gestion des connaissances - EGC-2009, Strasbourg France", 2009, p. 479-481, http://hal.inria.fr/inria-00436670/en/.

[52] B. FUCHS, A. NAPOLI. *Éléments de réflexion sur les composants d'ontologies et leur manipulation par RàPC*, in "Actes du 17ième séminaire sur le raisonnement à partir de Cas, Paris", B. FUCHS, A. NAPOLI (editors), LORIA, Nancy, 2009, p. 107–113.

[53] 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.

[54] M. KAYTOUE, S. DUPLESSIS, A. NAPOLI. *L'analyse formelle de concepts pour l'extraction de connaissances dans les données d'expression de gènes*, in "Extraction et gestion des connaissances (EGC 2009), Strasbourg", J.-G. GANASCIA, P. GANÇARSKI (editors), Revue des Nouvelles Technologies de l'Information, vol. RNTI-E-15, Cépaduès-Éditions, 2009, p. 439–440.

[55] F. LE BER, J. LIEBER, A. NAPOLI. *Utilisation d'une algèbre temporelle pour la représentation et l'adaptation de recettes de cuisine*, in "17ème Séminaire Raisonnement à partir de Cas, Paris France", 2009, p. 141-149, http://hal.archives-ouvertes.fr/hal-00409087/en/.

[56] J. LIEBER. *Le moteur de raisonnement à partir de cas de WikiTaaable*, in "17ème atelier sur le raisonnement à partir de cas - RàPC-09, Paris France", B. FUCHS, A. NAPOLI (editors), 06 2009, http://hal.archives-ouvertes.fr/inria-00437337/en/.

Workshops without Proceedings

[57] F. ANGEVIN, E. KLEIN, J.-F. MARI, F. LE BER, K. ADAMCZYK, H. MONOD, C. LAVIGNE. *Relative impacts of closest fields and background pollen on GM adventitious presence rates in non-GM harvests*, in "Genetically Modified Crops Coexistence Conference (GMCC'09), Melbourne, Australia", 2009, p. 1–7, articles en ligne.

[58] Z. ASSAGHIR, P. GIRARDIN, A. NAPOLI. *Fuzzy logic approach to represent and propagate imprecision in agri-environmental indicator assessment*, in "IFSA EUSFLAT World Congress 2009, Lisbon, Portugal", 2009.

[59] J. COJAN, J. LIEBER. *Une approche de l'adaptation en raisonnement à partir de cas fondée sur l'optimisation sous contraintes*, in "Journées d'Intelligence Artificielle Fondamentale, Marseille France", L. CHOLVY, S. KONIECZNY (editors), 10 2009, http://hal.inria.fr/inria-00425030/en/.

[60] M. KAYTOUE, A. NAPOLI. *Classification de données numériques par treillis de concepts et structures de patrons*, in "Journées d'Intelligence Artificielle Fondamentale – IAF 2009, Marseille, France", L. CHOLVY, S. KONIECZNY (editors), 2009.

Scientific Books (or Scientific Book chapters)

[61] H. ASTUDILLO, V. CODOCEDO, G. CANALS, D. TORRES, A. DIAZ, A. NAPOLI, A. GOMES, M. PIMENTEL. *Combining Knowledge Discovery, Ontologies, Annotations, and Semantic Wikis*, in "Webmedia Minicourse Book (WebMedia 2009, Brazilian Symposium on Multimedia and the Web)", M. TEIXEIRA (editor), Brazilian Computer Society (SBC), 2009, http://hal.inria.fr/inria-00435659/en/.

[62] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *A Conformity Measure using Background Knowledge for Association Rules: Application to Text Mining*, in "Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction", Y. ZHAO, C. ZHANG, L. CAO (editors), IGI Global, 2009, http://hal.archives-ouvertes.fr/inria-00437237/en/.

[63] M.-D. DEVIGNES, M. SMAÏL-TABBONE. *Maîtriser les ressources numériques : biologie "in silico"*, in "Biologie L'ère numérique", M. ROUX (editor), CNRS Editions, 06 2009, p. 189-222, http://hal.inria.fr/inria-00433477/en/, Chapitre 7.

Books or Proceedings Editing

[64] B. FUCHS, A. NAPOLI (editors). *Actes du 17ième séminaire sur le raisonnement à partir de Cas, Paris (29–30 juin)*, LORIA, Nancy, 2009.

References in notes

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

[66] 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.

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

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

[69] 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.

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

[71] 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.

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

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

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

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

[76] B. FUCHS, J. LIEBER, A. MILLE, A. NAPOLI. *An Algorithm for Adaptation in Case-based Reasoning*, in "Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin", W. HORN (editor), IOS Press, Amsterdam, 2000, p. 45–49.

[77] B. GANTER, S. 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.

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

[79] P. GIRARDIN, C. BOCKSTALLER, H. V. DER WERF. *Assessment of potential impacts of agricultural practices on the environment the AGRO* ECO method*, in "Environmental Impact Assessment Review", vol. 20, n^{o} 2, 2000, p. 227–239.

[80] M. KAYTOUE-UBERALL, S. DUPLESSIS, A. NAPOLI. *Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes*, in "Modelling, Computation and Optimization in Information Systems and Management Sciences, Second International Conference, MCO 2008, Metz, France - Luxembourg, September 8-10, 2008. Proceedings", H. L. THI, P. BOUVRY, T. P. DINH (editors), Communications in Computer and Information Science 14, Springer, 2008, p. 439–449.

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

[82] 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.

[83] 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", vol. 5, 1997, p. 22 – 25.

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

[85] 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.

[86] 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", vol. 5, n^{o} 8, 1997, p. 1093–1108.

[87] C. PESQUITA, D. FARIA, A. O. FALCÃO, P. LORD, F. M. COUTO. *Semantic Similarity in Biomedical Ontologies*, in "PLoS Comput Biol", vol. 5, 2009.

[88] V. I. PÉREZ-NUENO, D. W. RITCHIE, J. I. 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", vol. 48, n^{o} 11, 2008, p. 2146–2165.

[89] V. I. PÉREZ-NUENO, D. W. RITCHIE, O. RABAL, R. PASCUAL, J. I. 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", vol. 48, n^{o} 3, 2008, p. 509–533.

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

[91] 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. KUZNETSOV, S. SCHMIDT (editors), LNAI 4390, Springer, Berlin, 2007, p. 51–65.

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

[93] L. SZATHMARY. *Symbolic Data Mining Methods with the Coron Platform*, Université Henri Poincaré (Nancy 1), 2006, Thèse d'informatique.

[94] 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.

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

[96] 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.

[97] M. D'AQUIN, C. BOUTHIER, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Knowledge Edition and Maintenance Tools for a Semantic Portal in Oncology*, in "International Journal on Human–Computer Studies", vol. 62, n^{o} 5, 2005, p. 619–638.