INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team Orpailleur

Knowledge Discovery guided by Domain Knowledge

Nancy - Grand Est

THEME SYM

Activity Report

2008

Table of contents

Orpailleur is a project-team at LORIA since te 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. Indeed, such application domains are of first importance for the working systems developed by the team.

1. Team

Research Scientist

Amedeo Napoli [Researcher (DR CNRS), HdR]
Marie-Dominique Devignes [Researcher (CR CNRS), HdR]
Bernard Maigret [Researcher (DR CNRS), HdR]
Yannick Toussaint [Researcher (CR INRIA)]

Faculty Member

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)]
Malika Smaïl-Tabbone [Associate Professor (MdC Université Henri Poincaré Nancy 1)]

Technical Staff

Florent Marcuola [Engineer]
Thomas Meilender [Engineer]

PhD Student

Zainab Assaghir [PhD Student (INRA Grant)]
Yasmine Assess [PhD Student (INCa Grant)]
Fadi Badra [PhD Student (MERT Grant)]
Alexandre Beautrait [PhD Student (ARC Grant) until February 2008, 15th]
Sid-Ahmed Benabderrahmane [PhD Student (INCa Grant)]
Rokia Bendaoud [PhD Student (ATER)]
Matthieu Chavent [PhD Student (CNRS-Région Grant)]
Julien Cojan [PhD Student (AMX Grant)]
Adrien Coulet [PhD Student (CIFRE contract), leaving after Thesis defense on October 10th 2008]
Léo Ghemtio [PhD Student (ANR Contract)]
Nicolas Jay [PhD Student and lecturer (Faculté de Médecine, UHP Nancy 1), Thesis defended on October 7th 2008]
Mehdi Kaytoue [PhD Student (MERT Grant)]
Nizar Messaï [PhD Student (ATER)]
Frédéric Pennerath [PhD Student and lecturer (Supélec Metz)]
Naziha Benamrouche [Visiting Student]

Post-Doctoral Fellow

Vincent Leroux [Post-Doctoral fellow (INCa Grant)]
Mohamed Rouane-Hacene [Post-Doctoral fellow (ANR Project)]

Visiting Scientist

Sergei Kuznetsov [Professor (High School of Economics, Moscow, Russia, July 2008)]
Dave Ritchie [Professor (Aberdeen University, April and May 2008)]

Administrative Assistant

Emmanuelle Deschamps [Secretary]

Other

Nada Mimouni [Master Student (from February until August 2008)]
Mathias Vantieghem [Master Student (from February until August 2008)]

2. Overall Objectives

2.1. Introduction

Knowledge discovery in databases –hereafter KDD– consists in processing a huge 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 (in the same way, KDD is compared with archeology [76]). 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, 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 enlarging 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 year, the Taaable project has been designed for participating in a challenge, namely the "Computer Cooking Contest"[1] which was held during the European Conference on Case-Based Reasoning (ECCBR) in September 2008 in Trier (Germany). This participation 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 (see §5.5 and §6.3.3). The Taaable system has been designed during the project (see http://taaable.fr), in collaboration with SILEX team of LIRIS Lyon (an already long-term collaboration) and RCLP team of LIPN Paris (see §7.2.5) For this first participation, the Taaable system has won the second place (out of nine systems) and has been declared as European Vice-Champion of the "Computer Cooking Contest" [28]. Many people in the team are already involved in a second round and hope to win the first place for the forthcoming year.

3. Scientific Foundations

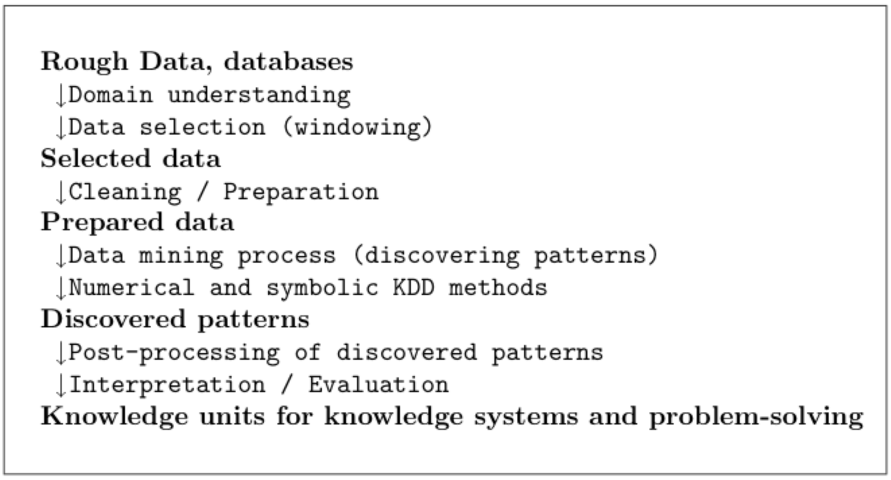
3.1. From KDD to KDDK

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

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 (see Figure 1).

[1] http://www.wi2.uni-trier.de/eccbr08/index.php?task=ccc



```
Rough Data, databases
 ↓Domain understanding
 ↓Data selection (windowing)
Selected data
 ↓Cleaning / Preparation
Prepared data
 ↓Data mining process (discovering patterns)
 ↓Numerical and symbolic KDD methods
Discovered patterns
 ↓Post-processing of discovered patterns
 ↓Interpretation / Evaluation
Knowledge units for knowledge systems and problem-solving
```

Figure 1. From data to knowledge units: the objective of the knowledge discovery process is to select, prepare and extract knowledge units from different data sources. For effective reuse, the extracted knowledge units have to be represented within an adequate knowledge representation formalism.

The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* [94], [95], [86] 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 [91]), frequent itemsets search, and association rule extraction [110].
- Numerical methods based on second-order Hidden Markov Models (HMM2, initially designed for pattern recognition [109], [108]). Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data.

Starting from these methods, the principle summarizing KDDK can be read as follows [42]: 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 [115], [81], [118], 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 [68], [70], involving text mining, content-based document mining, and intelligent information retrieval [77].

3.2. Symbolic Methods in Knowledge Discovery guided by Domain Knowledge

Keywords: *association rule extraction*, *formal concept analysis*, *frequent itemset search*, *knowledge discovery in databases guided by domain knowledge*, *lattice-based classification*.

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 [72], [91], [77]. 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). 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 [91].

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 [51], [50].

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.

3.3. Elements on Text Mining

Keywords: *document annotation*, *information extraction*, *knowledge discovery form large collection of texts*, *ontologies*, *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 [97], [81], [80]. The text mining process shows some 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. The interpretation of a text relies on knowledge units shared by the authors and the readers. A part of these knowledge units is expressed in the texts and may be extracted by the text mining process. Another part of these knowledge units, background knowledge, is not explicitly expressed in the text and is useful to relate notions present in a text, to guide and to help the text mining process. Background knowledge is encoded within an ontology (a knowledge base) associated to 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 such as frequent itemset search and association rule extraction [33], [32], [31]. 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, Semantic Web, and Web Mining

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

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.

Today 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. Tomorrow, the web will be "semantic" in the sense that people will search for information with the help of machines, that will be in charge of posing questions, searching for answers, classifying and interpreting the answers. The web will become 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 [66]). 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 [89]. Moreover, the adaptation step is deeply studied in the team (see for example [4]).

In the framework of semantic web, the mining of textual documents on the web, or web mining [74], can be considered from two main points of view: (i) mining the content of documents, involving text mining, (ii) mining the internal and external –hypertext links– structure of pages, involving information extraction. Web document mining is a major technique for the semi-automatic design of real-scale ontologies, the backbone of semantic web. In turn, ontologies are used for annotating the documents, enhancing document retrieval and document mining. In this way, web document mining improves annotation, retrieval, and the understandability of documents, with respect to their structure and their content. The extracted knowledge units can then be used for completing domain ontologies, that, in turn, guide text mining, and so on.

3.5. Data Mining with Hidden Markov Models

Keywords: *numerical data mining method*, *second-order Hidden Markov Models*, *stochastic process*.

A Hidden Markov Model is a stochastic process aimed at extracting and modeling a stationary distribution of events.

For designing a complete knowledge discovery system, we have developed stochastic models based on high-order hidden Markov models, namely second-order Hidden Markov Models (HMM2) for mining temporal and spatial data [62]. Hidden Markov Models have good capabilities to locate stationary segments (as shown in research work on speech recognition [108]). These models map sequences of data into a Markov chain in which transitions between states depend on the $\mathtt{n}$ previous states according to the order of the model ($\mathtt{n} = 2$ for HMM2). Actually, a second-order Hidden Markov model is defined as follows: (i) a set $\mathtt{S} = (\mathtt{s}_1, ...\mathtt{s}_\mathtt{N})$ of $\mathtt{N}$ states, (ii) a three dimensional matrix on $\mathtt{S}^3$ with $\mathtt{a}_{\mathtt{ijk}} = \mathtt{Prob}(\mathtt{q}_\mathtt{t} = \mathtt{s}_\mathtt{k}/\mathtt{q}_{\mathtt{t}-1} = \mathtt{s}_\mathtt{j}, \mathtt{q}_{\mathtt{t}-2} = \mathtt{s}_\mathtt{i})$, where $\mathtt{q}_\mathtt{t}$ denotes the state at time $\mathtt{t}$ and $\sum_{\mathtt{k}=1}^{\mathtt{N}} \mathtt{a}_{\mathtt{ijk}} = 1, \forall(\mathtt{i},\mathtt{j}) \in [1,\mathtt{N}] \times [1,\mathtt{N}]$, (iii) a set of $\mathtt{N}$ discrete distributions: $\mathtt{b}_\mathtt{i}(.)$ denotes for $\mathtt{i}$ the distribution of observations associated to the state $\mathtt{s}_\mathtt{i}$. This distribution may be parametric, non parametric, or even given by another Hidden Markov Model.

One research direction holds on the investigation of the performances of discrete second-order Hidden Markov Models on composite data, both in the temporal and spatial domain, to achieve a classification based on several attributes. The main advantage of HMM2 is the existence of a non-supervised training algorithm –the EM algorithm–, that allows the estimation of the parameters of the Markov model from a corpus of observations and an initial model. The resulting Markov model is able to segment each sequence of data into stationary and transient parts.

Our research effort focuses on the study of the capabilities of discrete second-order Hidden Markov Models applied to composite data, both in the temporal and spatial domain, to carry on a classification based on several attributes. This includes:

1. the elaboration of a process for mining spatial and temporal dependencies in order to extract knowledge units, e.g. for knowledge acquisition. This process involves an unsupervised classification of data whose results can be data processed by symbolic data mining methods (such as concept lattices or association rules).
2. The use of adapted symbolic classification methods, e.g. concept lattice design or association rules extraction algorithms, providing a bootstrap for automatic reasoning about the class structures extracted with the HMM2 and giving a synthetic view of the data to the experts, who have, in sequence, to interpret the classes and/or specify new experiments.

4. Application Domains

4.1. KDDK in Life Sciences

Keywords: *bioinformatics*, *biology*, *chemistry*, *gene*, *knowledge discovery in life sciences*.

Participants: Yasmine Assess, Sid-Ahmed Benabderrahmane, Naziha Benamrouche, Matthieu Chavent, Adrien Coulet, 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. Life sciences are getting more and more importance as an application domain for computer scientists. In this context, the understanding of biological systems provides complex problems for computer scientists. In sequence, when problems are solved (at least in part), solutions bring new ideas not only for biologists but also for computer scientists, and the research work goes on. Moreover, and this is one of the particularity of Orpailleur, the team includes biologists, chemists, and a physician, making the Orpailleur team one of the most original INRIA team from this point of view.

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

Keywords: *case-based reasoning*, *classification-based reasoning*, *description logics*, *knowledge representation*, *lattice-based classification*, *semantic web*.

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 [25]. This is a multidisciplinary research project in which participate researchers in computer science (Orpailleur), in ergonomics (“Laboratoire d’ergonomie du CNAM Paris”), 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 [88]. 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.

In parallel, the semantic web technology relies on the availability of large amount of knowledge in various forms [68], [70]. The acquisition of ontologies is one of the important issues that is widely explored in the semantic web community. Moreover, the acquisition of decision and adaptation knowledge for the semantic web 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 [120], that includes an OWL-based formalisms for representing medical ontologies, decision protocols (the case base), and adaptation knowledge [29]. Web services associated to the CBR process are developed, and several protocols are implemented.

4.3. Mining Spatio-Temporal Data

Keywords: *hidden Markov models*, *knowledge representation*, *spatial relations*, *spatial-temporal reasoning*.

Participants: Nicolas Jay, Florence Le Ber, Jean-François Mari, Amedeo Napoli.

Temporal and spatial data are complex data to be mined because of their internal structure, that can be considered as multi-dimensional. Indeed, spatial data may involve two or three dimensions for determining a region and complex relations as well for describing the relative positions of regions between each others (as in the RCC-8 theory for example [111]). Temporal data may present a linear but also a two-dimensional aspect, when time intervals are taken into account and have to be analyzed (using Allen relations for example). In this way, mining temporal or spatial data are tasks related to KDDK. Spatial and temporal data may be analyzed with numerical methods such as Hidden Markov Models, but also with symbolic methods, such as levelwise search for frequent sequential or spatial patterns.

For illustration, an application in the medical domain in concern with the study of chronic diseases is a good example of KDDK process on spatio-temporal data. An experiment for characterizing the patient pathway using the extraction of frequent patterns, sequential and not sequential, from the data of the PMSI[2] system associated with the "Lorraine Region" is currently under investigation [98].

5. Software

5.1. A Data Mining Toolkit: the Coron Platform

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

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

One of the goals of data mining is to extract hidden relations among objects and properties in databases. Usually frequent itemsets are used to find association rules, but the process produces a large number of rules, leading to the associated problem of "mining the set of extracted rules". Studies have shown that it can be more interesting to find only a subset of frequent itemsets, namely *frequent closed itemsets* (FCIs) and *frequent generators* (FGs). In turn, FCIs and FGs can be used for finding "minimal non-redundant" association rules.

We have developed a collection of programs for data mining that are grouped in the so-called Coron platform [116]. The platform contains a rich set of well-known algorithms in the data mining community, such as APriori, APriori-Close, Close, Pascal, Eclat, Charm, and, as well, several original algorithms such as Pascal+, ZART, Carpathia, Eclat-Z, and Charm-MFI. The toolkit is composed of three main parts: (i) Coron-base, (ii) AssRuleX, and (iii) pre- and post-processing modules.

With Coron-base, it is possible to extract different kinds of itemsets, e.g. frequent itemsets, frequent closed itemsets, frequent generators, etc. Each of the algorithms has advantages and disadvantages with respect to the form of the data that are mined. Since there is no best universal algorithm for any arbitrary dataset, Coron-base offers the possibility for users to choose the algorithm that best suits their dataset and needs.

Finding association rules is one of the most important tasks in data mining. The second part of the system, AssRuleX (Association Rule eXtractor) can generate different sets of association rules (from itemsets). This can lead to another data mining problem: which rules are the most useful? Beside all possible rules, some useful rule subsets (bases) can be extracted, e.g. minimal non-redundant association rules, generic basis, informative basis, etc.

The Coron system supports the whole life-cycle of a data mining task. We have 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.

[2] For "Programme de Médicalisation des Systèmes d'Informations". This is the name of the information system collecting the administrative data for an hospital.

Until now, studies in data mining have mainly concentrated on frequent itemsets and generation of association rules from them. Recently, we started to investigate the complement of frequent itemsets, namely the rare (or non-frequent) itemsets. In the literature, the problem of rare itemset mining and the generation of rare association rules has not yet been studied in detail, though such itemsets also contain important information just as frequent itemsets do. A particularly relevant field for rare itemsets is medical diagnosis. Coron already contains some algorithms that are designed to extract rare itemsets and rare association rules, e.g. APriori-rare, MRG-EXP, ARIMA, and BTB [116].

The Coron toolkit is developed entirely in Java, which provides a maximal portability. The system is operational, and it has already been tested 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). Moreover, the Coron implementation of the Titanic algorithm has been integrated into the Galicia platform, that is developed at the University of Québec (UQAM) in Montréal, Canada. Another extension of the system, named BioCoron, is aimed at taking into account biological data [56], [41].

The software has been registered at the “Agence pour la Protection des Programmes” (APP) and is freely available[3].

5.2. Stochastic systems for knowledge discovery and simulation

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

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

5.2.1. *CarottAge*

CarottAge [4] is a freely available software (GPL license) providing a synthetic representation of temporal and spatial data based on a hidden Markov field. The purpose of the CarottAge system is to build a partition –called the hidden partition– in which the inherent noise of the data is withdrawn as much as possible.

The software has been improved for taking into account:

1. the various shapes of the territories that are not represented by square matrices of pixels,
2. the use of pixels of different size with composite attributes representing the agricultural pieces and their attributes,
3. the irregular neighborhood relation between those pixels,
4. 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].

5.2.2. *GenExp*

In the framework of the project “Impact des OGM” initiated by the French Ministry of Research, we have developed a software called GenExp for simulating two-dimensional random landscapes and for studying the dissemination of vegetable transgenes. The GenExp system is based on the CarottAge system and on computational geometry. The simulated landscapes are given as input for programs such as `Mapod-Maïs` or `GeneSys-Colza` for studying the transgene diffusion [17], [63], [26]. The last version of GenExp allows an interaction with R subroutines and has received a GPL License.

5.3. Softwares for the Semantic Web

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

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

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

[5] See Project Piren-Seine, Inedit 65 “STIC and Environment”.

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

5.3.1. IntoWeb: Intelligent Access to Information

Two systems are currently under development. The objective of the first system, called "IntoWeb", is to provide a generic environment for an intelligent access to information, by combining information retrieval, hypertext navigation, and data-mining. Two kinds of objects feed the IntoWeb system: XML documents of a domain (for example bibliographical references) and web textual documents. The IntoWeb system contains a set of operations implementing the core tasks of a knowledge extraction process, i.e. collecting, filtering, and mining data. Applying operations to an object or a set of objects produce new objects, like vectors, clusters, association rules, lattices, which can in turn be exploited. Solving a given problem of information retrieval, or data mining, is performed by a well chosen sequence of operations available in the hypertext interface of the system. Data-mining modules, such as extraction of frequent closed itemsets, association rules, and lattice construction, are provided by the Coron platform (see §5.1).

5.3.2. CreChainDo

The second system, called "CreChainDo", makes use of FCA for information retrieval on the web. Many recent systems use FCA for improving the access to documents on the web [77], [99], [85]. Among them, the Credo system [78], 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 [21], [58]. 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. Thus, the CRECHAINDO system proposes an interactive and iterative information retrieval process on the web and is available[6].

5.4. 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 tools have been developed within the Kasimir system: mainly modules for the 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.

The software CabamakA for case base mining for adaptation knowledge acquisition is a module of the Kasimir system.

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, has to be carried out. A service of CBR based on an OWL representation has been developed for this purpose (see the thesis of Mathieu d'Aquin [121], [123]).

[6] http://intoweb.loria.fr/

The current release of EdHibou is stable and its use does not require the knowledge of complex web technologies, besides the web ontology language OWL or an OWL editor, such as Protégé+OWL. Moreover, a tutorial for EdHibou users has been written. Another tutorial, for EdHibou developers, is under writing. EdHibou has been presented at the International Semantic Web Conference, in the "poster and demos" program [29].

5.5. 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, Rokia Bendaoud, 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, manual and semi-automatic ontology engineering, text-mining from web textual resources, text annotation, knowledge representation, and hierarchical classification [28]. Ontologies for representing knowledge about the cooking domain, and a terminological base for binding texts and ontology concepts, have been semi-automatically built from textual web resources (see §6.2.1). These resources are used by an annotation process for building a formal representation of textual recipes (see §6.2.2). A CBR engine (see §6.3.3) 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 of the first "Computer Cooking Contest"[7] which took place during the European Conference on Case-Based Reasoning (ECCBR) in September 2008 in Trier (Germany).

6. New Results

6.1. Relational Concept Analysis and Mining Complex data

Keywords: *mining of complex data*, *relational concept analysis*.

Participants: Rokia Bendaoud, Nicolas Jay, Amedeo Napoli, Frédéric Pennerath, Mohamed Rouane-Hacene, Yannick Toussaint.

6.1.1. Relational Concept Analysis

Lattice-based classification, formal concept analysis, itemset search and association rule extraction, are suitable paradigms [117] for symbolic 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 [101], and on adaptability, i.e. the ability to fit evolving situations with respect to the constraints that may be associated with the KDDK process. Accordingly, the research work presented hereafter is in concern with the extension of symbolic methods to complex data, e.g. objects with multi-valued attributes, relations, graphs, texts, and real world data.

[7] http://www.wi2.uni-trier.de/eccbr08/index.php?task=ccc

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 adaptations of the classical FCA methodology for handling contexts holding on complex data formats, e.g. graph-based or relational data. This year, we have worked and developed several applications involving relational concept analysis (RCA) [55], [49]. The RCA process is an extension of the FCA process for analyzing objects described both by binary and relational attributes. 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. Moreover, a way of representing the concepts and relations extracted with RCA is proposed in the framework of a description logic. The RCA process has been implemented within the Galicia platform, offering new and efficient tools for knowledge and software engineering. Hereafter, several applications in text mining are currently using the RCA process and have given substantial results, showing the efficiency of the RCA process.

6.1.2. KDDK in Medico-Economical Databases

Since 30 years, many patient classification systems (PCS) have been developed. These systems aim at classifying care episodes into groups according to different patient characteristics. In most PCS, patient categories are derived from diagnoses and treatment procedures to get homogeneous groups in relation to resource use. This process is achieved by variance analysis of a numeric measure of resource consumption, explained by expert-defined patient profiles. Besides, the majority of PCS are designed to treat only a single encounter of a patient with a care provider at a time. In France, the so-called “Programme de Médicalisation des Systèmes d’Information” is a national wide PCS in use in every hospital. Each year, it collects data about millions of hospitalizations. Our objective is to extract new knowledge units from this database for exploring Patient Care Trajectories (PCT). PCT can be seen as sequences of several episodes of care observed over time. To perform this task, we propose a methodology based on Formal Concept Analysis (FCA). Our approach aims at assisting domain experts with automated classification tools to define groups of patient having similar health condition, treatments or journey through the health-care system.

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. These techniques are based on two interest measures of formal concepts: support and stability. Significant results have been be presented this year, especially at the International Conference Formal Concept Analysis 2008 [39], [40], [3]. They show the complementarity of these measures in identifying interesting concepts. Another way of research lies in the design of a data driven ontology. The idea is to reuse knowledge discovered during the FCA step to extend an ontology of PCT that will draw 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 [100] 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 aimed at discovering general synthesis methods from currently available chemical reaction databases for designing generic and reusable synthesis plans.

A preliminary research work [73] 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. This work has given substantial results for the expert chemists. At the moment, extending this first work, a graph-mining process is 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 is in accordance with needs of chemical industry [113]. This year, a number of substantial results have been presented (and some of these in high-level conferences [59], [60], [47], [48]) (see also 7.2.4).

6.2. KDDK and Text Mining

Keywords: *annotation*, *content-based manipulation of documents*, *document analysis and mining*, *ontology design and extension from texts*, *text mining*.

Participants: Rokia Bendoaud, Amedeo Napoli, Emmanuel Nauer, Mohamed Rouane-Hacene, Yannick Toussaint.

The objective of text mining is to extract new, 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. The results of research on text mining carried out during this year are based on two interrelated points: knowledge extraction from heterogeneous textual resources and semantic annotation.

6.2.1. Knowledge extraction from heterogeneous textual resources

Ontologies are the backbone of semantic web. They 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 [106]. An ontology is commonly defined as an explicit specification of a domain conceptualization [93]. However, in practice, building an ontology depends on a number of resources having different types: thesaurus, dictionaries, texts, databases, etc. The web makes also an increasing number of ontologies widely available for reuse. None of them can pretend to be complete as each ontology provides a specific point of view on a given domain. In addition, ontologies should be most of the time linked to each other. Then, because of this wide range of resources, it cannot be assumed that all ontologies are represented within a unique standard ontology language such as OWL, the W3C web Ontology Language (as sometimes assumed, e.g. in the framework of semantic import of modular ontologies [112]).

During this year, we have worked on the design of a methodology and an associated platform named "Pactole" [49], [33]. This methodology extends previous research works based on FCA and aimed at building ontologies from texts with Relational Concept Analysis (RCA). Accordingly, experiments have been performed mainly in two domains, astronomy and microbiology. Finally another experiment has been carried out in the Taaable project on the adaptation of cooking recipes.

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), information extraction tools (e.g. Gate), or any other ad hoc parsing tool. Properties are expressed in texts with adjectives or verbs (a celestial object is "flaring"). Relations are usually expressed through lexical patterns.

Based on the results of the extraction process, a binary table "Objects × Attributes" is built and the associated concept lattice can be computed. Objects are then clustered into classes w.r.t. the properties associated to the objects in the texts. Moreover, a transformation function may convert the lattice into a concept hierarchy expressed in a simple description logic formalism (FLE). Apposition of formal contexts (i.e. a merge of two contexts having the same set of objects) has been used for extending and enriching the resulting ontology. The RCA process has been used to take into account relations between objects and to create relation between concepts of the ontology.

Two experiments in astronomy and microbiology have been used for positively evaluating this approach [33], [32]. An interactive process based on FCA and RCA has also been defined, for inserting the user/expert/analyst into the KDD loop when building an ontology [33].

The same approach has been used in the Taaable project for building the Taaable ontology. A preliminary hierarchy of ingredients was taken from the Cook's Thesaurus[8]. Then, the possible ways of preparing an ingredient, e.g. mashed, sliced, etc., were used to refine the initial classification and to make closer ingredients that can be prepared the same way. In this way, an ontology of 4.000 ingredients was built and used by the case-base reasoner in the Taaable project.

6.2.2. Semantic Annotation of Cooking Recipes

A specific tool for annotating recipes w.r.t. an ontology has been developed in the context of the Taaable project. In the recipes, ingredients are identified thanks to a terminological matching mechanism. One problem is to annotate recipes by "types" (e.g. soup, dessert, Mediterranean dish, etc., since these types are not available in the initial recipe texts. Based on the experience of the team in symbolic KDD, association rule extraction has been applied on a set of 87.000 recipes. This resulted in the learning of significant features able to characterize recipe types. Accordingly, a set of selected association rules has been used for typing recipes.

6.2.3. KDDK in Pharmacovigilance

Participants: Amedeo Napoli, Mohamed Rouane-Hacene, Yannick Toussaint.

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.

During 2008, we participate to the ANR project Vigitermes, whose primary goal is to design a knowledge-based system for the management and the documentation of case reports, and, as well, to the detection of unexpected pharmacological associations.

We proposed a new method identifying candidates for pharmacological associations that can be investigated in clinical trials. A clinical trial allows the observation of a drug activity on a given population. The method relies on Formal Concept Analysis [91], a mathematical framework for data analysis. In addition, the proposed method uses several components of the statistical modeling and adjustments that help filtering statistically significant associations. The method has been implemented within a prototype system called SignalMiner and validated through an experiment on the CRDB of the French Medicines Agency. Moreover, a research paper introducing the method, the tool and the results has been submitted to Journal of Pharmaceutical Medicine.

In order to document and explain discovered pharmacological associations, we reformulated drug-reaction analysis so that extracting knowledge in CRDBs is mapped to a relational data-mining problem. Additional medical resources are used, containing taxonomic knowledge and information on both drugs and ADR. To that end, we rely on Relational Concept Analysis (RCA) [96], [114], an extended FCA framework that helps abstracting concepts and inter-concept relations based on patients, drugs, ADR, and their connections (see also §6.1.1). The goal here is to derive classes of patients (profiles) that are linked to, on the one hand, drug classes (therapies), and, on the other hand, reaction classes. The classes of drugs and reactions are provided with a formal description encoding chemical and pharmacokinetic data, medical classifications, etc. This helps to better detect the root causes of reactions in the light of the connections between patient profile and the mined knowledge on the followed therapy. The approach has been implemented within the tool Radix and is still under investigation in collaboration with Vigitermes stakeholders.

[8] http://www.foodsubs.com/

6.3. New directions within the Kasimir research project

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

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

6.3.1. Adaptation Knowledge Acquisition

The research about adaptation within the Kasimir research project has been described in [20]. Adaptation in Kasimir, as well as in many other CBR systems, requires knowledge. The adaptation knowledge acquisition (AKA) is a current research work, that takes two directions: AKA from experts and semi-automatic AKA.

AKA from experts consists in analyzing adaptations performed by experts. Interviews of experts confronted to decision problems requiring adaptation have been recorded to be afterward analyzed, and modeled within adaptation patterns [122].

Semi-automatic AKA is based on the "mining of the protocols". A protocol can be seen as a set of rules "`situation`⟶`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 has been implemented thanks to a frequent close itemset extraction module of the Coron platform (see §5.1). This requires a formatting of the situations and decisions of the protocol following the itemset mode. A system, called CabamakA, realizes this case base mining for adaptation knowledge acquisition, and provides pieces of information that can be used for building adaptation rules [119]. This AKA process is not fully automated: an analyst guides CabamakA, following the principles of knowledge discovery. More precisely, the analyst uses filters to drive the mining process, and interprets the extracted pieces of information in adaptation rules. The work of the analyst may be tedious, and thus tools making the task easier are under study and development [30].

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

6.3.2. Conservative Adaptation and Learning from Failure

A new scientific theme has emerged last year that can be applied to protocol adaptation. It is called *conservative adaptation* [34], [53] and it is a new approach to adaptation in CBR based on the theory of belief revision [71], [104], [105]. Revising a knowledge base ψ by a consistent knowledge base μ consists in building a knowledge base $\psi \circ \mu$ that specializes μ and that makes a "minimal change" on ψ in order to reach the consistency. Conservative adaptation consists in making a "minimal change" on the source context ψ (for Kasimir, the protocol) in order to be consistent with the target context μ (for Kasimir, the target patient). For example, suppose that the protocol recommends a cure of tamoxifen (an anti-oestrogen drug) for a patient having a cirrhosis of the liver. Since tamoxifen is contraindicated for patients having a serious liver disease, conservative adaptation proposes another anti-oestrogen treatment that is not contraindicated for this patient. In the continuity of this work, an approach involving a combination of several retrieved cases for solving a given target problem has been proposed last year [103].

6.3.3. CBR in the Taaable Project

Some of the results of the work on CBR in the Orpailleur team, in particular, for Kasimir, have been reused for the Taaable system (see also §5.5). This system is based on a hierarchy of classes where a class plays the role of recipe index. The retrieval in Taaable is based on strong and smooth classifications [102]. The result of retrieval is a recipe index and a similarity path in the form of a sequence of generalization/specialization substitutions. The adaptation consists in applying this sequence of substitutions on the index. Several research issues are currently examined in the Taaable project. One of them is the application of CabamakA on the recipes for finding cooking domain-dependent adaptation rules. Another one is the application of conservative adaptation for adapting recipes, taking into account domain knowledge about cooking. A long term future work consists in combining the rule-based adaptation –with rules obtained thanks to CabamakA– and conservative adaptation.

Finally, as they are collaborators of first importance in the Taaable project, let us mention a long term collaboration about the foundations of the CBR paradigm existing between Orpailleur and the SILEX team of the LIRIS laboratory in Lyon (see also §7.2.5). In 2008, this collaboration has focused on the IakA approach to knowledge acquisition of CBR [35]. This approach is opportunistic and interactive: it is applied in the context of a particular problem-solving session, when the user is not satisfied by the solution provided by the CBR system. It is based on a decomposition of the adaptation process: the user criticizes these steps and pieces of knowledge are acquired in a conversational manner, so that (1) the same failure does not occur again, and (2) the whole knowledge base of the CBR system improves. The context of this research was the thesis of Amélie Cordier [82].

6.4. KDDK in Life Sciences

Keywords: *bioinformatics*, *biology*, *chemistry*, *gene*, *knowledge discovery in life sciences*.

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

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 huge volumes of data would ultimately 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.

6.4.1. Model-driven Data Integration for Gene Retrieval (MODIM)

Our experience shows that the first step (data preparation) of the KDDK process is particularly complex and time consuming when dealing with biological data. We propose to rationalize this step by adopting a relevant methodology of MOdel-driven Data Integration for Mining (MODIM) [83]. This methodology can be summarized in 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 the previous work on an Approach for Candidate Gene Retrieval (ACGR)[9]. It has been applied for the IP3L project (see below). In both cases the data model lead to a relational database integrating relevant data. Alternatively, the same principles guided the construction of an integrated knowledge base for pharmacogenomics (see below).

6.4.2. Construction and use of a pharmacogenomic knowledge base for data integration and knowledge discovery

This work concentrates on ontology-guided data preparation for mining in the pharmacogenomics domain. The goal of pharmacogenomics is to discover knowledge about interactions between clinical, genetic and therapeutic data items. We studied the use of an ontology and a knowledge base for guiding the different steps of the KDDK process in the domain of pharmacogenomics [2]. Data related to this domain are heterogeneous, complex, and distributed over several data sources. Consequently, the preliminary step consisting in the preparation and the integration of data is crucial. For guiding this step, an original approach is proposed, based on a knowledge representation of the domain within two ontologies represented in description logics: namely SNP-Ontology and SO-Pharm.

[9] A paper has been recently accepted in the “Bioinformatics” journal.

SO-Pharm and SNP-Ontology are available in the OBO Foundry[10]. This approach has been implemented using semantic web technologies and has been used for populating a pharmacogenomic knowledge base. The “data ” to analyze are no more rough data but are objects represented within a knowledge base, making this approach very original. In addition, this approach is also beneficial for guiding operations in the knowledge discovery process. Firstly, we studied this benefit for feature selection by illustrating how the knowledge base can be used for this purpose [14]. Secondly, we described a new method named *Role Assertion Analysis* (or RAA) that enables knowledge discovery directly from a knowledge base [37], instead of a database. This method uses data mining algorithms over assertions (expressions represented within a description logics and involving individuals) of the pharmacogenomic knowledge base. This process has provided substantial results in the discovery of new and relevant knowledge.

6.4.3. Knowledge-oriented Filters for Improving High-throughput 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 previously reported VSM-G software [6], [5], [1], [9], [10], [11], [12], [15], [16], [23], [24]. 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), which is a target of interest. We therefore set up a KDDK approach on a dataset relative to LXR known active/inactive ligands. The first step consisted in integrating heterogeneous data pertaining from either public data sources or specific software following MODIM methodology. This resulted in the P3LI database (Protein-Protein and Protein-Ligand Interactions) which was designed ton encompass various mining tasks on proteins, ligands, and known protein-ligand and protein-protein interactions. A first mining scenario was performed for extracting association rules characterizing physical-chemical properties of active LXR ligands. The best retrieved rules are currently being evaluated as a filtering step [54].

6.4.4. Classifying Biological Databases with multi-valued FCA

Classification of biological DBs is crucial for facilitating their discovery and exploitation. The BioRegistry project aims at organizing metadata about biological DBs in order to prepare classification and retrieval tasks. Automatic procedures were designed and implemented to build a metadata repository starting from a curated web catalog [38]. Whenever possible, metadata are valued using existing domain ontologies (such as the MeSH thesaurus).

Formal Concept Analysis (FCA) was set up for organizing the BioRegistry DBs according to shared metadata. A formal context representing the relation between DBs and their metadata is exported from the repository, and the corresponding concept lattice is built [27]. In the basic setting of FCA, a formal context has the form of a binary table with rows corresponding to objects and columns corresponding to attributes. Since all metadata types do not have the same importance, we defined a first extension of FCA by introducing dependencies on attributes leading to the definition of attribute hierarchies which are considered for concept construction [45]. Complexity of DB metadata lead us to propose a second extension of FCA for 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 basic idea is that two objects share an attribute if the values taken by this attribute for these objects are similar, i.e. if the similarity is higher than a given threshold [46]. The present results are very promising and now we are trying to extend and generalize the approach, trying to set on a general extension of FCA in the same way as RCA does.

[10] http://www.obofoundry.org/cgi-bin/detail.cgi?id=pharmacogenomics

6.4.5. *Knowledge Discovery from Transcriptomic Data*

Transcriptomic data are produced by high-throughput devices and provide a quantitative measurement of the distribution of gene transcripts in a biological sample. Analyzing such data for a given set of samples, representing defined biological situations (cells grown under various conditions, various tissues, various stages of development of a given organ, disease stages, etc.), leads to define expression profiles. KDDK methodology can be advantageously applied for interpreting such profiles in terms of gene functions and regulatory pathways. In collaboration with Olivier Poch at IGBMC (Strasbourg) we have obtained in 2007 an INCa (Institut National du Cancer) funding for a PhD thesis. The work concerns the interpretation of transcriptomic data from colo-rectal cancer samples. Statistical pre-processing of the data has been performed in the Strasbourg laboratory. The first exploration of the provided datasets (220 genes, 3 situations) consisted in two steps. Firstly, we proposed a fuzzy modeling of differential gene expression between two situations. A consequence of this modeling is for example that given two situations and a threshold, some genes may simultaneously belong to over-expressed and equally expressed groups of genes. Fuzzy computation of differential expressions lead us to identify nine different expression profiles in our dataset. One tenth of the studied genes belongs to more than one profile. In a second step, we applied FCA as a preliminary attempt to interpret the expression profiles. A formal context was built in which each gene is described by its expression profile(s) and by functional annotations retrieved from public databases where genes are annotated using a controlled vocabulary called Gene Ontology (GO). Several concepts were thus constructed, revealing which functional features are shared by genes sharing the same expression profile. More experiments are now required to further investigate the biological meaning of each expression profile. The success and limits of major related approaches have to be summarized through a careful analysis of the abundant literature of the domain [87], [75]. Because the problem is clearly multi-dimensional, we intend to consider relational data mining techniques [67]. An ongoing work applies relational concept analysis (RCA) methods [55] to complete the results of FCA study by taking account known interactions between genes. We plan to validate the mining results by using another relational data mining method.

6.4.6. *Analyzing Groups of Co-Expressed with FCA*

This research work involves biologist from INRA (Champenoux) working on the associations between mushrooms and trees. In this research work, we study gene expression data (GED) [56], [41]. Microarray biotechnology is able to quantitatively measure the expression of a gene in a given biological environment or situation, which is relative to its activity. In this way, a so-called gene expression profile (GEP) can be considered as a numerical m-dimensional vector, describing the behavior of the gene. Then, gene expression data (GED) is a collection of n gene expression profiles that may be represented as an $n \times m$ numerical table, where each line is the profile of a gene. FCA is the method chosen for analyzing and interpreting the data, knowing that a widely admitted hypothesis states that genes having a similar expression profile may participate in a same biological function or process. In this way, formal concepts in the resulting concept lattice are representing sets of genes that present similar quantitative variations of expression in certain biological situations or environments. FCA is used both for its capabilities in data-mining, data organization, and knowledge representation. The data mining operations are articulated around three steps: numerical data is turned into binary data, then formal concepts are extracted and filtered thanks to the introduction of domain knowledge. This method has been successfully applied to a real dataset related to a fungus, named “Laccaria Bicolor”, for studying interaction between fungus and poplars (a very important tree in the industry of wood).

6.5. KDDK and Spatial Reasoning

Keywords: *lattice-based classification of relations*, *spatial and temporal reasoning*.

Participants: Zainab Assaghir, Florence Le Ber, Jean-François Mari, Amedeo Napoli.

In this framework, we work on the design of reasoning models on the representation of spatial structures in knowledge-based systems, e.g. hierarchical classification and CBR. This research work is applied to answer agronomic questions on the recognition and the analysis of farmland spatial structures. There are also ongoing studies on the design of concept lattices for mining and understanding complex hydrobiological data [7]. 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.

6.5.1. CBR on Spatial Organization Graphs

This work has been initiated in the framework of a thesis (2000–2005) done in collaboration with INRA SAD. The objective was to develop a knowledge-based system, called Rosa, for comparing and analyzing farm spatial structures. Reasoning in Rosa follows the principles of case-based reasoning (CBR) and relies on the agronomic assumption that there exists a strong relation between spatial and functional organizations of farms, i.e. similar spatial organizations correspond to similar functional organizations. According to this assumption, and given a set of previously studied farm cases, the Rosa system tries to help agronomists to analyze new problems holding on land use and land management in farms. Moreover, analysis of knowledge acquisition and modeling processes is continuing with the collaboration of researchers in socio-psychology and linguistics (Codisant, LabPsyLor, Université Nancy 2 and ICAR UMR 5191 CNRS, Lyon) [19], [18], [22], [8], [57].

6.5.2. Modeling Design Episodes

This work has been undertaken within the COPT ANR-ADD project (Conception d'Observatoires de Pratiques Territorialisées) in collaboration with CEVH, ENGEES-Université Louis Pasteur in Strasbourg, and Codisant, LabPsyLor, Université Nancy 2. The focus is on the experience of researchers in charge of building "*observatoires*", i.e. information systems for the monitoring and the management of rural territories. The goal here is to build a system taking into account past experiences for helping these researchers. We rely on previous work on experience-based reasoning [107], story-telling [69] and computer-aided design [79].

Actually, several persons in charge of "*observatoires*" have been interviewed and the resulting audio corpus has been mined in order to extract "design episodes", that are formalized parts of the design process. A database and a web interface have been developed to allow the consultation of these design episodes.

6.5.3. The definition of indicators in agronomy

To help farmers to improve their agricultural practices, researchers from French National Institute for Agronomic Research (INRA) in Colmar have proposed a set of agro-ecological *indicators*. Actually, an indicator estimates the impact of cultivation practices on the agrosystem and its environment [92]. The modeling and the assessment of environmental risk generally require a large number of parameters whose measure is imprecise. In this research work, we focus on various types of imprecision to be combined for defining the error associated with the measure given by indicator. The imprecision is introduced either by the estimations of the coefficients of the indicator or by measurement errors. In order to control this imprecision, the input variables of an indicator can be represented by fuzzy sets or possibilities distributions [84]. Then, based on this choice, the propagation of imprecision in the polymorphic calculus process of an indicator can be computed and controlled [52], [61].

6.6. Mining with HMMs: Results and Applications

Keywords: *agronomy*, *numerical data mining with hidden Markov models*.

Participants: Florence Le Ber, Jean-François Mari.

Several applications have been carried out during this year. In the project called "BiodivAgrim" holding on the domain of "Biodiversité", we are associated with the UR 55 of INRA SAD (Mirecourt). This ANR project has been launched in November 2007 and is the successor of the ACI ECOGER project (for "Écologie pour la Gestion des Écosystemes et de leurs Ressources") which has been extended for one extra year from 2008 to 2009.

The ADD COPT project, for “Agriculture et Développement Durable”, is running in its last year. Our work in the Piren Seine project took part in this project. It consisted in building a spatial partition of the Seine watershed (roughly 200 000 Km^2) based on the land cover successions. This partition is currently compared to those coming from other more conventional data analysis methods and to administrative regions. All this research work has taken advantage of the CarottAge system, a generic data-mining system for spatio-temporal data, based on HMM2 (see §5.2.1).

6.6.1. Two Applications in Bioinformatics

In the framework of the so-called “Contrat de Plan État-Région”, we are carrying out a long term data mining project with the laboratory of genetics and microbiology of the “Université Henri Poincaré Nancy 1” in two different areas: the detection of promoters (regulation motifs) and understanding of horizontal transfer.

We are currently studying a new data mining method based on stochastic analysis and combinatorial methods for discovering new transcriptional factors in bacterial genome sequences. Sigma factor binding sites (SFBS) are described as patterns of box1–spacer–box2 corresponding to the -35 and -10 DNA motifs of bacterial promoters. We are using a high-order Hidden Markov Model in which the hidden process is a second-order Markov chain. Applied to the genome of the model bacterium *Streptomyces coelicolor A3(2)*, the a posteriori state probabilities revealed local maxima or peaks whose distribution was enriched in the intergenic sequences (“iPeaks” for intergenic peaks). Short DNA sequences underlying the iPeaks (called “iPeak-motifs”) were extracted and clustered by a hierarchical classification algorithm based on the SmithWaterman local similarity. Some selected motif consensuses were used as box1 (-35 motif) in the search of a potential neighboring box2 (-10 motif) using a word enumeration algorithm. This new SFBS mining methodology applied on Streptomyces coelicolor was successfully used to retrieve already known SFBS and to suggest new potential transcriptional factor binding sites (TFBS). The well defined SigR regulon (oxidative stress response) was also used as a test quorum to compare first and second-order HMM. Our current approach also allows the preliminary detection of known SFBSs in *Bacillus subtilis*.

In the framework of the horizontal transfer understanding, we have validated the effectiveness of the HMM2 in the detection of exogenic areas and variable components within a genome of *Streptococcus Thermophilus*. The analysis of the genome of CNRZ1066 with Markov models of various orders has continued with the modification of the extraction parameters to optimize the extraction of the atypical areas. Finally, they are 146 atypical areas extracted, which account for 17.7% of the genome. The interpretation of part of these atypical areas was carried out by a comparative study with the genome LMD-9 recently published and by carrying out a comparison with 47 sequences of *Streptococcus Thermophilus* provided by a Danish company CHR. Hansen interested by this work.

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: Mehdi Kaytoue-Uberall, Amedeo Napoli, Yannick Toussaint.

The main goal of this cooperation project is define methodological and software support for integrating semantic features in a computer-supported design activity (models, methods and tools). The project intends to demonstrate that semantic web technologies are a suitable and efficient option for improving computer-assisted design process. For achieving this objective, semantic requirement for design situation have to be identified. Then, methodological support and software solution for management have to be developed.

The project has to support collaborative design activities guided by domain knowledge. Such an infrastructure has to use domain ontologies 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. Accordingly, the main research lines are knowledge representation and management for design activities and processes, and support to collaborative design activities. Moreover, a special attention will also be given to semantic wikis, in two way, semantic activities for wiki design and wiki management for semantic activities. In other word, one way is how to use ontologies to improve semantic wikis, and the other is how to use wikis for improving ontology design. The project is ending its first year and papers are on the way.

This project involves researchers form LORIA (Orpailleur and ECOO teams), LIFIA at Universita de La Plata (Argentina), Laboratório Intermidia at Universidade de São Paulo (Brazil), and Departamento de Informática at Universidad Técnica Federico Santa María (Chile).

7.1.2. Other international collaborations

Participants: Yasmina Assess, Naziha Benamrouche, Matthieu Chavent, Marie-Dominique Devignes, Mehdi Kaytoue-Uberall, Bernard Maigret, Amedeo Napoli, Mohamed Rouane-Hacene, Malika Smaïl-Tabbone, Yannick Toussaint.

We would like to mention to other close research international collaborations: (i) the first with Petko Valtchev at Université du Québec à Montréal (UQAM), and (ii) the second 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.

For the first collaboration, some research papers can already be mentioned [55], [49], [51], [50]. This research holds on the design of algorithms for itemset search and association rule extraction, and the design of concept lattices and relational concept lattices.

For the second collaboration, papers are submitted and the research topic is on the extension of FCA for taking into account complex structures. For that, we intend to extend and adapt previous work [90] and apply this extension to the analysis of complex biological data [56], [41].

Both collaborations are fruitful and we are trying to find an adequate context for funding in a better way the travels and the stays of the researchers.

Another important collaboration for the team is currently established with Dave Ritchie, Lecturer at Aberdeen University and creator of the docking program HEX, based on spherical harmonics and widely used. This collaboration was initiated in 2008 thanks to an INRIA visiting professor fellowship. Dave Ritchie visited the Orpailleur team from April 22nd until May 31st, 2008. There are at present already two outcomes resulting from this collaboration. Firstly, Dave Ritchie, Matthieu Chavent, and Bernard Maigret, obtained a very good rank at the CAPRI international challenge for prediction of 3D structure of a biomolecular complex. Secondly, the application involving Dave Ritchie for an “ANR Chaire d’Excellence” is successful. Thus, Dave Ritchie will join the Orpailleur team at the beginning of 2009 for two years, in order to develop innovative approaches for structural systems biology.

Finally, there is also an on-going exchange program related to an international collaboration with Algeria in the context of the so-called Tassili program (2006–2009).

7.2. National initiatives

7.2.1. ANR Nutrivigène

Participants: Florent Marcuola, Nada Mimouni, Amedeo Napoli, Mathias Vantieghem.

Nutrigenomics is an emerging topic whose interest is considerable for elaborating dietary recommendations and new food products. Homocysteine is an intermediate product of the one carbon metabolism that is closely correlated with age and associated with vascular, cognitive and neurological dysfunctions. It is also an end-product of trans-methylations of DNA and histones, reactions involved in epigenetic mechanisms of gene expression that depend in part on the status in folate and vitamin B12. The objective of the Nutrivigène project is to address the following question. At the cellular level, does the intermediate (50-100 micromol/L) and/or the moderate (15-50 micromol/L) hyperhomocysteinemia produce epigenetic changes of the expression of genes potentially related with the vascular, cognitive and neurological dysfunctions of the elderly? This study includes a mechanistic approach on a rat model deficient in dietary methyl precursors and an association study on the elderly volunteers of the cohort OASI.

The OASI cohort study has been designed to evaluate the association of gene-gene and gene-environment interactions with the risk for cardiovascular and neurological disorders in a rural region of Sicily. It will follow annually 3000 volunteers (door to door recruitment) from a rural mountain area (5 villages) of the north Sicily during 5 years. 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 aging. The evaluation of the association between the methylation of gene candidates (MS-PCR) and the vascular and cognitive function will be carried out by KDDK methods. At present, studies and investigations are still on the way, and papers are in preparation for describing the first analysis that have been carried on.

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: Mohamed Rouane-Hacene, Yannick Toussaint.

Pharmacovigilance covers research activities related to detection, analysis, and prevention of unexpected adverse drug reactions (ADR). In France, health-care professionals have the obligation to declare serious or unexpected ADRs. 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, the so-called “Agence Française de sécurité sanitaire des produits de santé”, maintains the database. Serious ADR cases are transmitted to the pharmaceutical laboratories that commercialize the associated drugs.

At the international level, the WHO (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 facilitating the work of the Pharmacovigilance staff –pharmaceutical industry and regulatory agencies– in its interaction with available resources, e.g. Pharmacovigilance database, summary of product characteristics, medical literature. The long term objective is to improve the process of signal detection in pharmacovigilance. Knowledge management in pharmacovigilance is based on formal representation of drugs and ADRs within medical terminologies.

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.

7.2.3. Research Projects in Computational Chemistry

Participants: Yasmina Assess, Naziha Benamrouche, Matthieu Chavent, Marie-Dominique Devignes, Léo Ghemtio, Vincent Leroux, Bernard Maigret, Malika Smaïl-Tabbone.

7.2.3.1. *FAK family tyrosine kinases: structural basis of regulation and localization (ANR Blanc 2005-2008).*

"Focal adhesion kinase" (FAK) forms an original group of non-receptor tyrosine kinases, which play an essential role in cell adherence, survival, and mobility, as well as in the development and function of various tissues including the nervous and immune systems. Kinases are involved in pathological conditions ranging from cancer invasion and metastasis to brain ischemia. The general goal of this project is to decipher the structural basis of the activation and cellular localization of FAK, and to design specific inhibitors of this enzyme. This project is ambitious and innovative, and needs a multidisciplinary approach. This is why experts in molecular, cellular, and structural biology, and as well, experts in molecular modeling and pharmacochemists are associated in this project. The combination of these complementary expertise is in favor of making significant progress in a research field, important for both fundamental science and therapeutic applications.

7.2.3.2. *Apelin: a potential therapeutic target for the treatment of heart failure: from the design of non peptidic agonists to clinical investigations (ANR cardiovasculaire 2005-2008) .*

Our efforts to clone an angiotensin III receptor led to the cloning of a gene encoding a rat seven transmembrane domain receptor exhibiting 95% sequence identity with that of the human orphan APJ receptor. This orphan has now been de-orphanized and corresponds to apelin receptor. The goal of this project is to investigate the structure-function relationships of apelin and its receptor, using structural (3D modeling) and pharmacological (hit discovery and optimization) approaches. In addition, a series of experiments has to be carried out for evaluating these molecules in vitro, for determining the effect of apelin on diuresis, and finally for determining whether apelin is involved in the control of water balance in humans.

7.2.3.3. *Synthetic and molecular modeling chemistry falls with biology for treatment of Met-triggered tumour growth and metastasis (projet non thématique INCa 2005-2008).*

This research project is at the cutting edge of cancer research as it aims at developing drugs based on novel chemical scaffold with therapeutic potential for Met-triggered cancer and metastatic processes. The driving force of our research is the multidisciplinary approaches we have undertaken to tackle a social problem like cancer, by integrating chemical synthesis, molecular modeling, structural and crystallography studies, optimization and functional validation in vitro and in vivo. Deregulated receptor tyrosine kinases (RTK) signals underlie unrestrained growth, metastatic behavior, and tumor angiogenesis. Specific RTK-based drugs are currently evaluated in clinical trials for tumor treatment. The proto-oncogene RTK Met plays a major role during malignant transformation of neoplastic cells. The objective of this project is to carry on the following tasks: (1) characterization and optimization of novel Met chemical inhibitors and their validation in animal models as suitable preclinical test for therapeutic strategies, (2) identification of chemical compounds able to block molecular pathways underlying metastatic processes. This research makes use of drug design based on synthetic and molecular modeling chemistry, tissue cultures studies, and genetically modified animals. The combination of these approaches should lead to methods for diagnosis, prevention, and treatment of this cancer.

7.2.4. Projets Exploratoires Pluridisciplinaires CNRS (PEPS)

A PEPS project was funded by the CNRS on the topic: “differential exploitation of transcriptomic data: semantic integration and mining”. This multidisciplinary project involves the IGBMC laboratory in Strasbourg and more precisely the group of Olivier Poch. This collaboration lead us to define a subject of PHD thesis and to obtain a grant from the INCa institute (Institut National du Cancer) via the Canceropole Grand-Est. This PhD thesis is co-directed by members ofv the Orpailleur team, M.-D. Devignes and M. Smail-Tabbone, and by Olivier Poch.

Another PEPS project holds on the mining of chemical reaction databases in chemistry. This work is done in collaboration with chemists in Montpellier and is aimed at setting on a methodology and a process of graph-mining for extracting synthesis methods from chemical reaction databases in chemistry. Synthesis methods can be seen as strategic tools for guiding a chemical synthesis. A paper is in preparation for summarizing the research work that has been carried out for the last two years.

7.2.5. Collaborations around the Taaable Project

The Taaable system (see also §5.5) was designed in collaboration with the SILEX team (LIRIS Lyon) and the RCLN team (LIPN Paris 13), for participating in the first “Computer Cooking Contest” (CCC)[11], which was organized during the European Conference on Case-Based Reasoning (ECCBR) in September 2008 in Trier (Germany). The participants to this challenge were invited to propose a knowledge-based system, e.g. a case-based reasoning system, for solving a set of cooking problems in using a given set of textual recipes. The Taaable system won the second price. The Taaable staff has decide to extend the present system, to design a second version, and to participate in the next challenge CCC to be held at the forthcoming 2009 ICCBR.

7.2.6. Projects and Collaborations in Spatio-Temporal Reasoning

- Programme fédérateur “Agriculture et Développement Durable”: Conception d’Observatoires de Pratiques Territorialisées de la Durabilité de l’Agriculture (COPTDA) (in charge of Jean-François Mari).
- Collaborations: ENGEES Strasbourg, INRA in Nancy-Mirecourt, Paris-Grignon, Dijon, and Toulouse, Laboratoire ESE UPRESA 8079 CNRS/Paris-Sud, Équipe Codisant, LPI GRC, Université de Nancy 2, GRIC UMR 5612 CNRS Lyon, and ENGREF Clermont-Ferrand.

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).

  A research operation is currently under development on knowledge management and decision support in the large within the Kasimir system.

[11] http://www.wi2.uni-trier.de/eccbr08/index.php?task=ccc

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, again 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

Year Publications

Doctoral Dissertations and Habilitation Theses

[1] A. Beautrait. *Développement et validation de la plate-forme de criblage virtuel VSM-G*, Thèse de chimie-informatique, Université Henri Poincaré (Nancy 1), 2008.

[2] A. Coulet. *Construction et utilisation d’une base de connaissances pharmacogénomique pour l’intégration de données et la découverte de connaissances*, Thèse d’informatique, Université Henri Poincaré (Nancy 1), 2008.

[3] N. Jay. *Découverte et représentation de trajectoires de soins par analyse formelle de concepts*, Thèse d’informatique, Université Henri Poincaré (Nancy 1), 2008.

[4] J. Lieber. *Contributions à la conception de systèmes de raisonnement à partir de cas*, Habilitation à diriger des recherches, spécialité informatique, Université Henri Poincaré (Nancy 1), 2008.

Articles in International Peer-Reviewed Journal

[5] A. Beautrait, A. Karaboga, M. Souchet, B. Maigret. *Cluster Induced fit in liver X receptor beta: a molecular dynamics-based investigation*, in "Proteins", vol. 72, 2008.

[6] A. Beautrait, V. Leroux, M. Chavent, L. Ghemtio, M. Devignes, M. Smail-Tabbone, W. Cai, X. Shao, G. Moreau, P. Bladon, J. Yao, B. Maigret. *Multiple-step virtual screening using VSM-G: overview and validation of fast geometrical matching enrichment.*, in "Journal of Molecular Modeling", vol. 14, 2008.

[7] A. BERTAUX, F. LE BER, A. BRAUD, M. TRÉMOLIÈRES. *Mining Complex Hydrobiological Data with Galois Lattices*, in "International Journal of Computing & Information Sciences", To be published., 2008.

[8] C. BRASSAC, S. LARDON, F. LE BER, L. MONDADA, P.-L. OSTY. *Analyse de l'émergence de connaissances au cours d'un processus collectif*, in "Revue d'Anthropologie des Connaissances", vol. 2, n^{o} 2, 2008, p. 267–289.

[9] W. CAI, J. XU, X. SHAO, V. LEROUX, B. MAIGRET. *SHEF: a vHTS geometrical filter using coefficients of spherical harmonic molecular surfaces*, in "Journal of Molecular Modeling", vol. 14, 2008.

[10] M. CHAVENT, B. LEVY, B. MAIGRET. *MetaMol: High-quality visualization of molecular skin surface.*, in "Journal of Molecular Graphics and Modeling", vol. 27, 2008.

[11] C. CLAPERON, R. ROZENFELD, X. ITURRIOZ, N. INGUIMBERT, M. OKADA, B. ROQUES, B. MAIGRET, C. LLORENS-CORTES. *Contribution of molecular modeling and site-directed mutagenesis to the identification of threonine 348 as a residue involved in aminopeptidase a substrate specificity.*, in "Journal of Biological Chemistry", 2008.

[12] C. CLAPERON, R. X. ROZENFELD, X. ITURRIOZ, N. INGUIMBERT, M. OKADA, B. ROQUES, B. MAIGRET, C. LLORENS-CORTES. *Asp-218 participates with asp-213 to bind a ca2+ atom into the S1 subsite of aminopeptidase A: a key element for substrate specificity.*, in "Biochemical Journal", vol. 416, 2008.

[13] A. COULET, M. SMAÏL-TABBONE, P. BENLIAN, A. NAPOLI, M.-D. DEVIGNES. *Ontology-guided data preparation for discovering genotype-phenotype relationships*, in "BMC Bioinformatics", vol. 9, n^{o} Suppl 4, 2008, S3.

[14] A. COULET, M. SMAIL-TABBONE, P. BENLIAN, A. NAPOLI, M.-D. DEVIGNES. *Ontology-guided data preparation for discovering genotype-phenotype relationships*, in "BMC Bioinformatics", vol. 9, 2008.

[15] 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", 2008.

[16] M. FOUCAUD, E. ARCHER-LAHLOU, E. MARCO, I. TIKHONOVA, B. MAIGRET. *Insights into the binding and activation sites of the receptors for cholecystokinin and gastrin.*, in "Regulatory Peptides", vol. 145, 2008.

[17] C. LAVIGNE, E. KLEIN, J.-F. MARI, F. LE BER, K. ADAMCZYK, H. MONOD, F. ANGEVIN. *How do genetically modified (GM) crops contribute to background levels of GM pollen in an agricultural landscape?*, in "Journal of Applied Ecology", vol. 45, n^{o} 4, 2008, p. 1104–1113.

[18] F. LE BER, C. BRASSAC. *Étude longitudinale d'une procédure de modélisation de connaissances en matière de gestion du territoire agricole*, in "Revue d'Anthropologie des Connaissances", vol. 2, n^{o} 2, 2008, p. 151–168.

[19] F. LE BER. *Introduction au dossier "De la bergerie au centre de calcul : élaboration collective de connaissances spatiales"*, in "Revue d'Anthropologie des Connaissances", vol. 2, n^{o} 2, 2008, p. 147–149.

[20] 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.

[21] E. NAUER, Y. TOUSSAINT. *CreChainDo: an iterative and interactive Web information retrieval system based on lattices*, in "International Journal of General Systems", 2008.

[22] P.-L. OSTY, F. LE BER, J. LIEBER. *Raisonnement à partir de cas et agronomie des territoires — Constructions croisées*, in "Revue d'Anthropologie des Connaissances", vol. 2, n^{o} 2, 2008, p. 169–193.

[23] S. PATANÉ, N. PIETRANCOSTA, H. HASSANI, V. LEROUX, B. MAIGRET, J. KRAUS, R. DONO, F. MAINA. *A new Met inhibitory-scaffold identified by a focused forward chemical biological screen.*, in "Biochemical and Biophysical Research Communications", vol. 375, 2008.

[24] A. TOUMI-MAOUCHE, B. MAOUCHE, S. TAIRI-KELLOU, S. EL-AOUFI, M. MARTIN-MARTINEZ, R. GONZALEZ-MUNIZ, D. FOURMY, B. MAIGRET. *Exploring the binding pocket for pyridopyrimidine ligands at the CCK1 receptor by molecular docking.*, in "Journal of Molecular Modeling", vol. 14, 2008.

[25] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Decision Support Systems in Oncology: Are we there yet?*, in "Oncology", vol. 9, n^{o} 11, 2008, http://www.hcplive.com/mdnglive/ONCNG-Oncology/Nov2008.

Articles in National Peer-Reviewed Journal

[26] K. ADAMCZYK, F. ANGEVIN, N. COLBACH, C. LAVIGNE, F. LE BER, J.-F. MARI. *GenExP, un logiciel simulateur de paysages agricoles pour l'étude de la diffusion de transgènes*, in "Revue Internationale de Géomatique", vol. 17, n^{o} 3-4, 2007, p. 469–487.

[27] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Correction et complétude d'un algorithme de recherche d'information par treillis de concepts*, in "Revue des Nouvelles Technologies de l'Information RNTI", Numéro spécial coordonné par M. Nadif et F.-X. Jollois, vol. Classification : points de vue croisés, 2008, p. 147–158.

International Peer-Reviewed Conference/Proceedings

[28] 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.

[29] F. BADRA, M. D'AQUIN, J. LIEBER, T. MEILENDER. *EdHibou: a Customizable Interface for Decision Support in a Semantic Portal*, in "International Semantic Web Conference (Posters & Demos)", 2008.

[30] F. BADRA, J. LIEBER. *Representing Case Variations for Learning General and Specific Adaptation Rules*, in "Fourth Starting AI Researchers' Symposium (STAIRS 2008)", 2008, p. 1–11.

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

[32] 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.

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

[34] J. COJAN, J. LIEBER. *Conservative Adaptation in Metric Spaces*, in "Advances in Case-Based Reasoning, 9th European Conference, ECCBR-2008, Trier, Germany. Proceedings", LNAI 5239, 2008, p. 135–149.

[35] A. CORDIER, B. FUCHS, L. LANA DE CARVALHO, J. LIEBER, A. MILLE. *Opportunistic Acquisition of Adaptation Knowledge and Cases – The IakA Approach*, in "Advances in Case-Based Reasoning, 9th European Conference, ECCBR-2008, Trier, Germany. Proceedings", LNAI 5239, 2008, p. 150–164.

[36] A. COULET, M. SMAÏL-TABBONE, A. NAPOLI, M.-D. DEVIGNES. *Ontology refinement through role assertion analysis: example in pharmacogenomics*, in "Proceedings of the 21st International Workshop on Description Logics (DL-2008), Dresden, Germany", F. BAADER, C. LUTZ, B. MOTIK (editors), CEUR Workshop Proceedings 353, CEUR-WS.org, 2008.

[37] A. COULET, M. SMAIL-TABBONE, A. NAPOLI, M.-D. DEVIGNES. *Ontology Refinement through Role Assertion Analysis: Example in Pharmacogenomics*, in "21st International Workshop on Description Logics - DL2008, Dresden Italie", F. BAADER, C. LUTZ, B. MOTIK (editors), 2008.

[38] M.-D. DEVIGNES, P. FRANIATTE, N. MESSAI, A. NAPOLI, M. SMAIL-TABBONE. *BioRegistry: Automatic Extraction of Metadata for Biological Database Retrieval and Discovery*, in "1st international workshop on Ressource Discovery (RED), Joint to iiWAS 2008, Linz Autriche", 2008.

[39] N. JAY, F. KOHLER, A. NAPOLI. *Analysis of social communities with iceberg and stability-based concept lattices*, in "Formal Concept Analysis - Proceedings of the 6th International Conference on FCA (ICFCA 2008), Montréal", R. MEDINA, S. OBIEDKOV (editors), Lecture Notes in Artificial Intelligence 4933, Springer, Berlin, 2008, p. 258–272.

[40] N. JAY, F. KOHLER, A. NAPOLI. *Using Formal Concept Analysis for mining and interpreting patient flows within a healthcare network*, 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. 263–268.

[41] 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.

[42] 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.

[43] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Extending Attribute Dependencies for Lattice-Based Querying and Navigation*, in "Proceedings of the 16th International Conference on Conceptual Structures, ICCS 2008, Toulouse, France", P. EKLUND, O. HAEMMERLÉ (editors), Lecture Notes in Computer Science 5113, 2008, p. 189–202.

[44] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Many-valued concept lattices for conceptual clustering and information retrieval*, in "18th European Conference on Artificial Intelligence (ECAI-08), Patras, Greece", M. GHALLAB, C. SPYROPOULOS, N. FAKOTAKIS, N. AVOURIS (editors), IOS Press, 2008, p. 127–131.

[45] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAIL-TABBONE. *Extending Attribute Dependencies for Lattice-Based Querying and Navigation*, in "16th International Conference on Conceptual Structures - ICCS 2008 Conceptual Structures: Knowledge Visualization and Reasoning Lecture Notes in Computer Science, Toulouse France", P. EKLUND, O. HAEMMERLÉ (editors), Lecture Notes in Computer Science, vol. 5113, Springer, 2008, p. 189–202.

[46] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval*, in "18th European Conference in Artificial Intelligence - ECAI 2008, Patras Grèce", M. GHALLAB (editor), IOS Press, 2008, p. 127–131.

[47] F. PENNERATH, G. POLAILLON, A. NAPOLI. *A Method for Classifying Vertices of Labeled Graphs Applied to Knowledge Discovery from Molecules*, in "Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-08), Patras, Greece", M. GHALLAB, C. SPYROPOULOS, N. FAKOTAKIS, N. AVOURIS (editors), IOS Press, 2008, p. 147–151.

[48] F. PENNERATH, G. POLAILLON, A. NAPOLI. *Mining Intervals of Graphs to Extract Characteristic Reaction Patterns*, in "Discovery Science", J.-F. BOULICAUT, M. BERTHOD, T. HORVÁTH (editors), Lecture Notes in Computer Science 5255, Springer, Berlin, 2008, p. 210–221.

[49] M. ROUANE-HACENE, A. NAPOLI, P. VALTCHEV, Y. TOUSSAINT, R. BENDAOUD. *Ontology Learning from Text using Relational Concept Analysis*, in "International Conference on eTechnologies (MCETECH 08), Montréal", P. KROPF, M. BENYOUCEF, H. MILI (editors), IEEE Computer Society, 2008, p. 154–163.

[50] L. SZATHMARY, P. VALTCHEV, A. NAPOLI, R. GODIN. *An Efficient Hybrid Algorithm for Mining Frequent Closures and Generators*, in "Proceedings of the Sixth International Conference on Concept Lattices and their Applications (CLA'08), Olomouc, Czech Republic", R. BELOHLAVEK, S. KUZNETSOV (editors), Palacký University, Olomouc, 2008, p. 47–58.

[51] 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.

National Peer-Reviewed Conference/Proceedings

[52] Z. ASSAGHIR, P. GIRARDIN, A. NAPOLI. *Utilisation de la théorie de possibilités pour calculer un indicateur environnemental*, in "Rencontres Francophones sur la Logique Floue et ses Applications (LFA), Lens (France)", Cépaduès Editions, 2008, p. 390–397.

[53] J. COJAN, J. LIEBER. *Formalisation de l'adaptation conservatrice dans les espaces métriques*, in "16ième atelier raisonnement à partir de cas (RàPC-2008)", 2008, p. 92–107.

[54] L. GHEMTIO, E. BRESSO, M. SOUCHET, B. MAIGRET, M. SMAÏL-TABBONE, M.-D. DEVIGNES. *Model-driven data integration for mining protein-ligand and protein-protein interactions in a drug design context.*, in "Journées Ouvertes Biologie Informatique Mathématiques (JOBIM), Lille France", 2008.

[55] M. HUCHARD, A. NAPOLI, M. ROUANE-HACENE, P. VALTCHEV. *Extraction de connaissances à partir de données relationnelles avec l'analyse formelle de concepts*, in "Conférence sur la reconnaissance des formes et l'intelligence artificielle (RFIA 2008), Amiens", P. MARQUIS, I. BLOCH (editors), AFRIF–AFIA, 2008, p. 143–152.

[56] M. KAYTOUE-UBERALL, S. DUPLESSIS, A. NAPOLI. *Toward the discovery of itemsets with significant variations in gene expression matrices*, in "Book of Short Papers - First Joint Meeting of SFC and CLADAG, Caserta (Italy)", B. FICHET, D. PICCOLO, R. VERDE (editors), Edizione Scientifiche Italiane, Napoli, 2008, p. 333–336.

[57] F. LE BER, C. BRASSAC. *Modéliser l'entre deux dans l'organisation spatiale des exploitations agricoles*, in "Journées Jean-Pierre Deffontaines, Versailles", INRA, 2008.

[58] E. NAUER, Y. TOUSSAINT. *Classification dynamique par treillis de concepts pour la recherche d'information sur le web.*, in "5ème Conférence de recherche en information et applications - CORIA 2008, Trégastel France", 2008, p. 71-86.

[59] F. PENNERATH, A. NAPOLI. *Le problème de l'extraction des graphes d'intérêt maximal – Application à la fouille de réactions chimiques*, in "Actes du 16ème Congrès Francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA 2008), Amiens", P. MARQUIS, I. BLOCH (editors), AFRIF–AFIA, 2008, p. 133–142.

[60] F. PENNERATH, G. POLAILLON, A. NAPOLI. *Prétraitement des bases de données de réactions chimiques pour la fouille de schémas de réactions*, in "Extraction et gestion des connaissances (EGC 2008), Sophia-Antipolis", RNTI, Cépaduès-Éditions Toulouse, 2008, p. 547–558.

Workshops without Proceedings

[61] Z. ASSAGHIR, P. GIRARDIN, A. NAPOLI. *Traitement d'imperfections des indicateurs agri-environnementaux*, in "2ème manifestation des jeunes chercheurs en STIC (MajecSTIC), Marseille", 2008.

[62] G. LAZRAK, M. BENOÎT, J.-F. MARI. *Arpentage: Analyse de Régularités Paysagëres pour l'Environnement dans les territoires Agricoles*, in "Symposium "Spatial landscape modelling: from dynamic approaches to functional evaluation", Toulouse", June 2008.

[63] F. LE BER, K. ADAMCZYK, C. LAVIGNE, J.-F. MARI. *GenExP, a software for simulating random field-patterns*, in "Spatial Landscape Modelling Symposium, Toulouse", 2008.

Scientific Books (or Scientific Book chapters)

[64] F. L. Ber, T. Libourel (editors). *Actes du colloque international "Spatial Analysis and GEOmatics" (SAGEO'08)- 25-27 juin 2008, Montpellier*, Actes sur CD, 2008.

Books or Proceedings Editing

[65] M. Bouzid, F. Le Ber, G. Ligozat, O. Papini (editors). *Actes du 4ème atelier Représentation et raisonnement sur le temps et l'espace (RTE 2008), SAGEO'08, Montpellier*, 2008.

References in notes

[66] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider (editors). *The Description Logic Handbook*, Cambridge University Press, Cambridge, UK, 2003.

[67] S. Dzeroski, N. Lavrac (editors). *Relational Data Mining*, Springer, Berlin, 2001.

[68] D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (editors). *Spinning the Semantic Web*, The MIT Press, Cambridge, Massachusetts, 2003.

[69] E. Soulier (editor). *Storytelling : Concepts, outils, applications*, Traité IC2, Hermès, Paris, 2006.

[70] S. Staab, R. Studer (editors). *Handbook on Ontologies*, Springer, Berlin, 2004.

[71] C. E. Alchourrón, P. Gärdenfors, D. Makinson. *On the Logic of Theory Change: partial meet functions for contraction and revision*, in "Journal of Symbolic Logic", vol. 50, 1985, p. 510–530.

[72] M. Barbut, B. Monjardet. *Ordre et classification – Algèbre et combinatoire (2 tomes)*, Hachette, Paris, 1970.

[73] S. Berasaluce, C. Laurenço, A. Napoli, G. Niel. *An Experiment on Knowledge Discovery in Chemical Databases*, in "Knowledge Discovery in Databases: PKDD 2004, Pisa", J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (editors), Lecture Notes in Artificial Intelligence 3202, Springer, Berlin, 2004, p. 39–51.

[74] B. Berendt, A. Hotho, G. Stumme. *Towards Semantic Web Mining*, in "The Semantic Web - ISWC 2002, Berlin", I. Horrocks, J. Hendler (editors), Lecture Notes in Artificial Intelligence 2342, Springer, 2002, p. 264–278.

[75] S. Blachon, R. Pensa, J. Besson, C. Robardet, J.-F. Boulicaut, O. Gandrillon. *Clustering formal concepts to discover biologically relevant knowledge from gene expression data*, in "In Silico Biology", vol. 7 (0033), 2007, p. 1–15.

[76] R. Brachman, P. Selfridge, L. Terveen, B. Altman, A. Borgida, F. Halper, T. Kirk, A. Lazar, D. McGuinness, L. Resnick. *Knowledge representation support for data archaeology*, in "Proceedings of the 1st International Conference on Information and Knowledge Management (CKIM'92), Baltimore", 1992, p. 457–464.

[77] C. Carpineto, G. Romano. *Concept Data Analysis: Theory and Applications*, John Wiley & Sons, Chichester, UK, 2004.

[78] 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.

[79] P.-A. Champin. *ARDECO: an assistant for experience reuse in Computer Aided Design*, in "Proceedings of WS 5 of ICCBR'03: From structured cases to unstructured problem solving episodes, Trondheim, Norvège", 2003, p. 287–294.

[80] H. Cherfi, A. Napoli, Y. Toussaint. *Towards a Text Mining Methodology Using Association Rules Extraction*, in "Soft Computing", vol. 10, n^o 5, 2006, p. 431–441.

[81] 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.

[82] A. Cordier. *Interactive and Opportunistic Knowledge Acquisition in Case-Based Reasoning*, Thèse d'informatique, Université Lyon 1, 2008.

[83] M.-D. Devignes, M. Smail-Tabbone. *Workshop PC Chairs' Message. Web Data Integration for Mining in the Life Sciences (WebDIM4LS)*, in "International Workshops on Web Information Systems Engineering WISE 2007 Workshops Web Information Systems Engineering WISE 2007 Workshops Lecture Notes in Computer Science, Nancy France", M. Weske, M.-S. Hacid, C. Godart (editors), Lecture Notes in Computer Science, vol. 4832, Springer, 2007, p. 3-4, http://hal.inria.fr/inria-00338679/en/.

[84] D. Dubois, H. Prade. *Théorie des possibilités : applications à la représentation des connaissances en informatique*, Masson, Paris, 1987.

[85] J. Ducrou. *DVDSleuth: A Case Study in Applied Formal Concept Analysis for Navigating Web Catalogs*, in "Conceptual Structures: Knowledge Architectures for Smart Applications, 15th International Conference on Conceptual Structures (ICCS 2007)", LNCS 4604, Springer, 2007, p. 496–500.

[86] M. Dunham. *Data Mining – Introductory and Advanced Topics*, Prentice Hall, Upper Saddle River, NJ, 2003.

[87] M. Eisen, P. Spellman, P. Brown, D. Botstein. *Cluster analysis and display of genome-wide expression patterns*, in "PNAS", vol. 95, 1998, p. 14863–14868.

[88] Evidence-based medicine working-group. *Evidence-based medicine. A new approach to teaching the practice of medicine*, in "Journal of the American Medical Association", vol. 17, 1992, 268.

[89] 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.

[90] 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.

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

[92] 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.

[93] T. GRUBER. *Toward principales for the design of ontologies used for knowledge sharing*, in "Formal Analysis in Conceptual Analysis and Knowledge Representation", N. GUARINO, R. POLI (editors), Kluwer Academic, 1993.

[94] J. HAN, M. KAMBER. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, 2001.

[95] D. HAND, H. MANNILA, P. SMYTH. *Principles of Data Mining*, The MIT Press, Cambridge (MA), 2001.

[96] M. HUCHARD, M. ROUANE-HACENE, C. ROUME, P. VALTCHEV. *Relational Concept Discovery in Structured Datasets*, in "Annals of Mathematics and Artificial Intelligence", vol. 49, n^{o} 1–4, 2007, p. 39–76.

[97] D. JANETZKO, H. CHERFI, R. KENNKE, A. NAPOLI, Y. TOUSSAINT. *Knowledge-based Selection of Association Rules for Text Mining*, in "16h European Conference on Artificial Intelligence – ECAI'04, Valencia, Spain", R. L. DE MÀNTARAS, L. SAITTA (editors), 2004, p. 485–489.

[98] N. JAY, F. KOHLER, A. NAPOLI. *Using Formal Concept Analysis for mining and interpreting patient flows within a healthcare network*, in "Fourth International Conference on Concept Lattices and their Applications (CLA-06), Hammamet, Tunisia", S. B. YAHIA, E. MEPHU-NGUIFO (editors), (See this volume), 2006.

[99] B. KOESTER. *Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results with Contexts and Concept Hierarchies*, in "Industrial Conference on Data Mining", Lecture Notes in Computer Science, vol. 4065, Springer, 2006, p. 176-190.

[100] S. KUZNETSOV. *Machine Learning and Formal Concept Analysis*, in "Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia", P. W. EKLUND (editor), Lecture Notes in Computer Science 2961, Springer, 2004, p. 287–312.

[101] S. KUZNETSOV, S. OBIEDKOV. *Comparing performance of algorithms for generating concept lattices*, in "Journal of Theoretical Artificial Intelligence", vol. 14, n^{o} 2/3, 2002, p. 189–216.

[102] J. LIEBER. *Strong, Fuzzy and Smooth Hierarchical Classification for Case-Based Problem Solving*, in "Proceedings of the 15th European Conference on Artificial Intelligence (ECAI-02), Lyon, France", F. VAN HARMELEN (editor), IOS Press, Amsterdam, 2002, p. 81–85.

[103] J. LIEBER. *Application de la révision et de la fusion des connaissances à l'adaptation et à la combinaison de cas*, in "Actes du quinzième atelier raisonnement à partir de cas, RàPC'07, Grenoble", A. CORDIER, B. FUCHS (editors), Plateforme AFIA, 2007, p. 119–129.

[104] J. LIEBER. *Application de la théorie de la révision à l'adaptation en raisonnement à partir de cas : l'adaptation conservatrice*, in "Actes des quatrièmes journées francophones sur les modèles formels de l'interaction", 2007, p. 201–213.

[105] J. LIEBER. *Application of the Revision Theory to Adaptation in Case-Based Reasoning: the Conservative Adaptation*, in "Proceedings of the 7th International Conference on Case-Based Reasoning, Belfast", Lecture Notes in Artificial Intelligence 4626, Springer, 2007, p. 239–253.

[106] A. MAEDCHE. *Ontologies Learning for the Semantic Web*, Springer, 2002.

[107] M. MAHER, A. G. DE SILVA GARZA. *Case-Based Reasoning in Design*, in "IEEE Expert", vol. 12, n^o 2, 1997, p. 34–41.

[108] 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.

[109] 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.

[110] 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.

[111] A. NAPOLI, F. LE BER. *The Galois lattice as a hierarchical structure for topological relations*, in "Annals of Mathematics and Artificial Intelligence", vol. 49, n^o 1–4, 2007, p. 171–190.

[112] J. PAN, L. SERAFINI, Y. ZHAO. *Semantic Import: An Approach for Partial Ontology Reuse*, in "1st International Workshop on Modular Ontologies (WoMO'06) In ISWC 2006", 2006.

[113] F. PENNERATH, A. NAPOLI. *La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique*, in "Extraction et gestion des connaissances (EGC'2006), Lille", G. RITSCHARD, C. DJERABA (editors), RNTI-E-6, Cépaduès-Éditions Toulouse, 2006, p. 517–528.

[114] 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.

[115] G. STUMME. *Formal Concept Analysis on Its Way from Mathematics to Computer Science*, in "Conceptual Structures: Integration and Interfaces, Proceedingsof the 10th International Conference on Conceptual Structures, ICCS 2002, Borovets, Bulgaria, Berlin", U. PRISS, D. CORBETT, G. ANGELOVA (editors), Lecture Notes in Artificial Intelligence 2393, Springer, 2002, p. 2–19.

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

[117] P. VALTCHEV, R. MISSAOUI, R. GODIN. *Formal Concept Analysis for Knowledge Discovery and Data Mining: The New Challenges*, in "Concept Lattices, Second International Conference on Formal Concept

Analysis, ICFCA 2004, Sydney, Australia", P. W. EKLUND (editor), Lecture Notes in Computer Science 2961, Springer, 2004, p. 352–371.

[118] R. WILLE. *Mathods of Conceptual Knowledge Processing*, in "International Conference on Formal Concept Analysis, ICFCA 2006, Dresden, Germany", R. MISSAOUI, J. SCHMID (editors), Lecture Notes in Artificial Intelligence 3874, Springer, 2006, p. 1–29.

[119] 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.

[120] 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.

[121] M. D'AQUIN. *Un portail sémantique pour la gestion des connaissances en cancérologie*, Thèse d'université, Université Henri Poincaré Nancy 1, soutenue le 15 décembre 2005, 2005.

[122] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Adaptation Knowledge Acquisition: a Case Study for Case-Based Decision Support in Oncology*, in "Computational Intelligence (an International Journal)", vol. 22, n^{o} 3/4, 2006, p. 161–176.

[123] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Case-Based Reasoning within Semantic Web Technologies*, in "Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2006), Varna, Bulgaria, 13-15th September, 2006", J. EUZENAT, J. DOMINGUE (editors), LNAI 4183, Springer, Berlin, 2006, p. 190–200.