INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team Orpailleur

Knowledge Discovery guided by Domain Knowledge

Nancy - Grand Est

THEME SYM

Activity Report

2007

Table of contents

1. Team

Head of project team

Amedeo Napoli [Researcher (DR CNRS), HdR]

Administrative assistant

Emanuelle Deschamps [Secretary]

Permanent researchers

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

Faculty members

Florence Le Ber [Professor (ENGEES Strasbourg), HdR]
Jean Lieber [Associate Professor (MdC Université Henri Poincaré Nancy 1)]
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

Charu Asthana [Engineer]
Florent Marcuola [Engineer, from September 2007, 15th]
Thomas Meilender [Engineer, from November 2007, 1st]

PhD Students

Yasmine Assess [PhD Student (INCa Grant)]
Fadi Badra [PhD Student (MERT Grant)]
Alexandre Beautrait [PhD Student (ARC Grant)]
Sid-Ahmed Benabderrahmane [PhD Student (INCa Grant), from October 1st]
Rokia Bendaoud [PhD Student (Région-INRIA Grant)]
Matthieu Chavent [PhD Student (CNRS-Région Grant)]
Julien Cojan [PhD Student (AMX Grant) from October 2007, 1st]
Adrien Coulet [PhD Student (CIFRE contract with Kika médical Nancy)]
Léo Ghemtio [PhD Student (ANR Contract)]
Nicolas Jay [PhD Student and lecturer (Faculté de Médecine, UHP Nancy 1)]
Mehdi Kaytoue [PhD Student (MERT Grant), from October 2007, 1st]
Mohamed Zied Maala [PhD Student (France Télécom Grant)]
Nizar Messaï [PhD Student (Région-UHP Nancy 1 Grant)]
Frédéric Pennerath [PhD Student and lecturer (Supélec Metz)]
Laszlo Szathmary [ATER at UHP Nancy 1)]
Sylvain Tenier [PhD Student (CIFRE contract with INIST Diffusion Nancy)]

Internships

Naziha Benamrouche [Visiting Student]
Bertrand Delecroix [Post-Doctoral fellow (until April 2007)]
Alexander Estacio Moreno [Post-Doctoral fellow (until June 2007)]
Vincent Leroux [Post-Doctoral fellow (INCa Grant)]
Mohamed Rouane Hacene [Post-Doctoral fellow (INRIA Grant)]

External collaborators

Sergei Kuznetsov [Professor (High School of Economics, Moscow, Russia, June 2007)]
Petko Valtchev [Associate Professor (UQAM Montréal, Québec, June and October 2007)]

2. Overall Objectives

2.1. Overall Objectives

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 [52]). 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!*

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).

The KDDK process –as implemented in the research work of the Orpailleur team– is based on *data mining methods* [64], [65], [60] 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 [63]), frequent itemsets search, and association rule extraction [76].
- Numerical methods based on second-order Hidden Markov Models (HMM2, initially designed for pattern recognition [74], [73]). 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: 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.

Table 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.

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		

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 [80], [58], [84], 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 [44], [46], involving text mining, content-based document mining, and intelligent information retrieval [54].

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 [49], [63], [54]. 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* [63] or FCA. Lattice-based classification is the basic operation underlying the so-called *formal concept analysis* [63] or FCA.

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 [40].

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) [41]. 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 [66], [58], [57]. 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 i.e. frequent itemset search and association rule extraction [22]. 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 [43]). 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 [62].

In the framework of Semantic Web, the mining of textual documents on the Web, or Web mining [51], 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*.

An 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 [15], [16]. Hidden Markov Models have good capabilities to locate stationary segments (as shown in research work on speech recognition [73]). These models map sequences of data into a Markov chain in which transitions between states depend on the n previous states according to the order of the model ($n = 2$ for HMM2). Actually, a second-order Hidden Markov model is defined as follows: (i) a set $S = (s_1, ...s_N)$ of N states, (ii) a 3 dimensional matrix on S^3 with $a_{ijk} = Prob(q_t = s_k/q_{t-1} = s_j, q_{t-2} = s_i)$, where q_t denotes the state at time t and $\sum_{k=1}^{N} a_{ijk} = 1, \forall(i, j) \in [1, N] \times [1, N]$, (iii) a set of N discrete distributions: $b_i(.)$ denotes for i the distribution of observations associated to the state s_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.

The research effort focuses on two main points: (1) the elaboration of a process for mining spatial and temporal dependencies in order to extract knowledge units (for knowledge acquisition). This process involves an unsupervised classification of data whose results will be the data processed by symbolic methods. (2) The specification of adapted symbolic classification tools giving a synthetic view of the data to the experts who have 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, Charu Asthana, Alexandre Beautrait, 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, Laszlo Szathmary, Yannick Toussaint.

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

The application domains that are currently investigated at the moment by the Orpailleur team are related with life sciences, with a particular emphasis on biology (bioinformatics), medicine, and chemistry as well. Indeed, there are various reasons explaining why life sciences are a major application domain. In general, life sciences are getting more and more importance as a domain application for computer scientists. In this context, the collaboration between biologists and computer scientists is very active, and the understanding of biological systems provides complex problems for computer scientists. When these problems are solved (at least in part), the solutions bring new ideas not only for biologists but also for computer scientists in their own research work. Thus, advances in research appear on both sides, life and computer sciences.

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: *Semantic Web*, *case-based reasoning*, *classification-based reasoning*, *description logics*, *knowledge representation*, *lattice-based classification*.

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

The KASIMIR research project holds on decision support and knowledge management for the treatment of cancer. 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).

The main tasks of the KASIMIR system are decision support and knowledge management for the treatment of cancer. 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 [61]. 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 such an out of the protocol case, oncologists try to *adapt* the protocol. Actually, considering the complex case of breast cancer, oncologists discuss such a case during the so-called “breast 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. Adaptation knowledge plays a key issue in applications, e.g. in knowledge-intensive case-based reasoning systems [47].

In parallel, the Semantic Web technology relies on the availability of large amount of knowledge in various forms [44], [46]. 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. A parallel research topic is to apply AKA to the extraction of decision knowledge units. Indeed, adaptation knowledge is closely related with decision theory, e.g. the Wald pessimistic criterion is frequently applied when pieces of information about a patient are missing. The AKA process is aimed at feeding a knowledge server embedded in the KASIMIR semantic portal [85], that includes an OWL-based formalisms for representing medical ontologies, decision protocols (the case base), and adaptation knowledge. Web services associated to the CBR process are developed. Several protocols are implemented, with a few of them including adaptation knowledge.

4.3. Mining Spatio-Temporal Data

Keywords: *hidden Markov models*, *knowledge representation*, *spatial relations*, *spatial-temporal reasoning*.
Participants: Charu Asthana, 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 [18]). 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[1] system associated with the “Lorraine Region” is currently under investigation [67].

5. Software

5.1. A Data Mining Toolkit: the Coron Platform

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

Participants: Mehdi Kaytoue, Florent Marcuola, Amedeo Napoli, Laszlo Szathmary [contact person], 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. 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 [40], 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. 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.

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 [41].

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.

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

The software has been registered at the "Agence pour la Protection des Programmes" (APP) and is freely available[2].

5.2. Stochastic systems for knowledge discovery and simulation

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

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

5.2.1. CarottAge

One aspect of data-mining is to provide a synthetic representation of data that a domain analyst can interpret. 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. Then spatio-temporal data are explored for extracting homogeneous classes both in temporal and spatial dimensions, giving also a clear view of the transitions between the classes.

CAROTTAGE is a free software, under a GPL license, taking as input an array of discrete data where the rows represent the spatial sites and the columns the time slots, and building a partition with the associated *a posteriori* probability. This probability may be plotted as a function of time, and is a meaningful feature for the analyst searching for stationary and transient behaviors of data. This software is currently used by INRA researchers interested in mining the successions of land use processes, e.g. in order to build models simulating the contamination of cave and surface waters.

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 bidimensional random landscapes, and then 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 [3]. This year, we have released a new version of GenExp allowing an interaction with R subroutines. This version is on the way to receive a GPL License.

5.3. Softwares for the Semantic Web

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

Participants: Mohamed Maala Zied, Amedeo Napoli, Emmanuel Nauer [contact person], Laszlo Szathmary, Yannick Toussaint.

5.3.1. IntoWeb: Intelligent Access to Information

Three 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 [33]. 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 hypertextual 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. DefineCrawler: a Generic Crawler

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

The "DefineCrawler" system can be seen as a Web information retrieval "meta-system", in the sense that the Web crawling can be parametrized. General parameters, like the maximum depth of the crawl, the set of starting URL, the number of parallel processes crawling the Web, the time allowed for crawling, etc. define the global behavior of the system. Validation parameters specify conditions –connected by Boolean operators– that must be satisfied by each document, for eliminating documents without interest with respect to the user need, e.g. documents that do not satisfy some criteria, that are not in a fixed language, etc. Evaluation parameters are additional conditions designed to evaluate the relevance of returned documents. Every evaluation and validation conditions is defined by an external instruction, allowing the use of various commands or tools, e.g. for checking the presence of an element, for counting the occurrences of some elements, for calculating a similarity between documents. The evaluation and validation conditions can be combined to calculate a score for a returned document. This score is then used to rank the returned documents.

5.3.3. CreChainDo

Another system (this is a new research work) is based the use of formal concept analysis (FCA) for information retrieval on the Web. Many recent systems use FCA for improving the access to documents on the Web [54], [69], [59] (see also [31]). Among them, the CREDO system [55], 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 [34]. 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[3].

5.4. The Kasimir System

Keywords: *case-based reasoning*, *classification-based reasoning*, *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 [86], [88]).

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

The current release of EDHIBOU has two main limitations. Firstly, it is rather difficult to be used by a naive user, not aware of a set of complex technologies related to Semantic Web. Secondly, the user-interface automatically generated from an ontology represented in OWL DL is not enough ergonomic: tools for a customization of this interface for a particular application have to be implemented.

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, Laszlo Szathmary, Yannick Toussaint.

6.1.1. Relational Concept Analysis

Lattice-based classification, formal concept analysis, itemset search and association rule extraction, are suitable paradigms [83] 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 [71], 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.

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) [9], [38], [39]. 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 have been currently using the RCA process and have yield 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 healthcare 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. First results will be presented at the International Conference Formal Concept Analysis 2008. 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 [70] 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 [50] 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 [77]. Substantial results will be presented in two forthcoming conference sin January of 2008 (see also 7.3.3).

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, Bertrand Delecroix, Alexander Estacio Moreno, Amedeo Napoli, Mohamed Rouane Hacene, Sylvain Tenier, Yannick Toussaint.

The objective of text mining is to extract new and useful knowledge units from large collections of texts, for allowing the machine-based manipulation of texts. The results of the research work carried out during this year are based on three strongly interrelated points: information extraction from texts, knowledge acquisition from texts, and semantic annotation.

6.2.1. Information Extraction from Texts

Information extraction from texts relies on natural language processing tools (NLP) for extracting information from texts. No NPL tool is ready to be immediately used on thousand of texts without a precise and thorough (and thus a time consuming) configuration. This year, we experimented the Information Extraction platform “Gate” and the “Link Grammar Parser” for English. Both are used to extract terminological entities from texts, their properties and their relations. We obtained the following results:

- Gate includes a rule-based system identifying in the texts units of information that are of interest for a user. After a 6-months training period holding on texts in bioinformatics, Gate has been configured for extracting relations between bacteria, antibiotics, and genes mutations. The relational concept analysis process has been applied to these data.

- The Link Grammar Parser was used on texts in astronomy and microbiology. >From these texts, pairs such as subject-verb, verb-object, and triples such as subject-verb-object, have been extracted and encoded for being processed by FCA and RCA tools.

6.2.2. Knowledge Acquisition from Texts

Following the current research work on FCA for building ontologies from texts [58], we extended this work to Relational Concept Analysis (RCA). An operational platform has been designed and evaluated in two application domains: astronomy and microbiology.

The main idea is that verbs may be used to characterize domain objects. In astronomy, verbs can be mainly associated to properties. For example, the sentence “We observed stars” makes stars “observable”. A binary table (Objects × Verbs) and the associated concept lattice have been built from texts. Objects are then structured into classes following the properties they are associated with in the texts. A transformation function converts the lattice into a concept hierarchy, where objects of the domain are terminal nodes.

A method based on apposition of lattices has been defined for extending and enriching an existing ontology from the the results of the FCA or RCA processes. In this way, an ontology is completed by the knowledge units that have been extracted from texts. An evaluation carried on by an expert in astronomy shows that new knowledge units have indeed been extracted from texts: identification of new celestial objects (objects that are not in the reference database), discovery of new properties for classes of objects, modification of the class of some astronomical objects. More precisely, the methodology and the prototype that are currently developed allows the extraction of significant elements as showed in the following:

- Dicovery of a new celestial objects: We identified in the texts the “protostellar cold cloud fragment HH 24MMS” which was recently discovered. In particular, the HH 24MMS celestial object was not known from the reference celestial object database (called the Simbad database, this work is done in association with researchers in astronomy, see http://cdsweb.u-strasbg.fr/).
- Some properties were previously unknown in the sense that no correlation between the celestial type of objects and the properties was known. For example “59 Aurigae” and “V1208 Aql” can pulse, “MM Herculis” can eclipse, or “AB Dor” and “OJ 287” can flare.
- A new class of objects has been proposed: its instances are “Orion” and “TWA”, with the property “expanding”. This new class has been called “Association of young stars” and it specializes the already known “Association of stars” class.

In another context, in microbiology, verbs are involved into relations much more than in in astronomy: “We have previously reported that a significant percentage (44%) of isoniazid-resistant Mycobacterium tuberculosis strains carry an arginine to leucine mutation in codon 463 (R463L) in the catalase-peroxidase gene (katG)”. The idea is then to organize bacteria, gene, and antibiotics, on the basis of the relations expressed in texts. Relational context analysis has been used for building classes of entities, properties of classes, and relations between classes.

Still in a close domain, a pharmacovigilance databases consist of several case reports involving drugs and adverse events (AEs). Some methods are applied consistently for detecting signals, i.e. statistically significant associations between a drug and an AE. These methods are appropriate for verification of complex relationships involving one or several drugs and AEs (e.g. syndromes or interactions) but do not address their identification. We have designed a method for extracting these kinds of relationships with FCA, in association with “disproportionality measures”. The method identifies all sets of drugs and AEs which are potential signals, syndromes or interactions. Compared to a previous experience of “disproportionality analysis” without FCA, the addition of FCA appears to be more efficient for identifying false positives related to concomitant drugs.

6.2.3. Semantic Annotation of Web Pages

Search engines should now be able to deal with complex queries that integrate user knowledge. In microbiology, for example, database queries about gene interactions return thousands of documents. In order to focus on a particular type of interaction, knowledge characterizing interactions is required. Querying web pages also implies to be able to reason on the content of these pages. Therefore, this content has to be explicitly represented within a formal language. According to the Semantic Web proposal [44], the content of a web page can be represented within a knowledge representation language, such as OWL DL. In that way, the content can be distributed, accessed and manipulated using knowledge-enabled software agents. The main results of this year are the development of a DL-based methodology for annotating Web pages. The main idea is to build a semantic representation of the page in accordance with the structure of the page. It is currently tested for scientific watch applied to the scientific research community.

6.3. New directions within the Kasimir research project

Keywords: *Semantic Web*, *case-based reasoning*, *classification-based reasoning*, *description logics*, *knowledge representation*, *lattice-based classification*.

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

6.3.1. Adaptation Knowledge Acquisition

The adaptation in KASIMIR, as well as in many 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 [87].

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 [42]. 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 [20], [21].

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 this year that can be applied to protocol adaptation. It is called *conservative adaptation* [29], [30] and it is a new approach to adaptation in CBR based on the theory of belief revision [48]. 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 in [28].

Conservative adaptation is based on domain knowledge. Following the previous example, domain knowledge must include the facts that a liver cirrhosis is a serious liver disease and that tamoxifen is contraindicated for such diseases. Without these knowledge units, conservative adaptation proposes tamoxifen. This result will be judged by an expert in oncology as inconsistent with domain knowledge. The prototype FRAKAS (FailuRe Analysis for domain Knowledge AcquiSition) interacts with the expert for identifying, on the basis of inconsistency, pieces of knowledge to be added to domain knowledge such that conservative adaptation with new domain knowledge does not produce inconsistency. FRAKAS was designed and developed in collaboration with researchers of LIRIS (a CNRS laboratory in Lyon) [24], [25].

The implementation of conservative adaptation and FRAKAS is based on a propositional knowledge representation language. This formalism is appropriate for validating the ideas, but is not enough expressive to represent most of the protocols in an appropriate way. A new study holds on the extension of conservative adaptation and of FRAKAS to description logics, which involves the necessity of defining revision operators in this knowledge representation formalism.

6.4. KDDK in Life Sciences

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

Participants: Yasmine Assess, Charu Asthana, Alexandre Beautrait, 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, Laszlo Szathmary, Yannick Toussaint.

Genome sequences, biological structures, expression arrays, proteomics, represent terabytes of data which are stored under variable formats in dispersed heterogeneous databases (DB). More than 800 such DBs have been listed at the beginning of 2006. One of the major challenges in the post genomic era consists in exploiting the vast amounts of biological data stored in those DBs. The extraction of knowledge from all these data is an increasingly challenging task which ultimately gives sense to the data production effort with respect to domains such as evolution and disease understanding, biotechnologies, system biology, or pharmacogenomics.

6.4.1. Model-driven Data Integration for Gene Retrieval

Understanding genetic diseases relies on the discovery of genetic defects responsible for the observed disorder. Computational methods are nowadays widely used to discover candidate genes by exploiting the mass of data accumulated in public genomic databases and eventually crossing these data with private data [53], [78], [81]. Usually, a similarity measure is computed between annotations of candidate genes and known disease genes. However the relation between a disease gene and a disease may be complex, especially in the case of rare diseases presenting mixed phenotypes. Exploiting prior knowledge about orthology relationships or interactions as well as data from multiple species is then required. We propose a novel approach for retrieving disease-specific candidate genes based on an integrated genomic information system involving three categories of candidate gene definitions: the first expresses direct relationship between a gene and a disease, the second category introduces intermediary genes (orthologuous or interacting genes), and the third category involves experimental data (such as transcriptomic data). These definitions were used for guiding data modeling and were converted into views which lead to the retrieval of sets of candidate genes. The implementation of the so-called ACGR approach uses a relational database management system. The data integration is driven by the data model and relies on the specification of several scenarios (encapsulated into wrappers). Three case-studies were performed on complex diseases including a rare disease for which responsible gene is still unknown. The user feedback and the results of the ACGR approach are very encouraging [79]. Results are reported in a paper currently submitted to the Bioinformatics journal.

6.4.2. KDDK in Pharmacogenomics

Another ongoing research work concentrates on ontology-guided data preparation for data mining in the domain of pharmacogenomics. The goal of pharmacogenomics is to discover knowledge about interactions between clinical, genetic, and therapeutic data.

Two formal ontologies, namely SNP-Ontology and So-Pharm, have been designed. SNP-Ontology formalizes knowledge about genomic variations and was used for integrating the various heterogeneous representations of both private data and data coming from public databases (dbSNP, UCSC, HapMap...). SO-Pharm ontology enlarges the SNP-Ontology scope for covering pharmacogenomics clinical trials, and, more precisely, for representing groups of individuals involved in trials, their genotype, their treatment, their observed phenotype, and the potential pharmacogenomics relations discovered between these concepts. The objective of SO-Pharm is to guide KDD in pharmacogenomics and was recently published in the OBO Foundry.

The three major steps of KDD could benefit from domain knowledge embedded in bio- ontologies. Several groups studied how ontologies can help the mining step [68] or the interpretation of the mining results [82]. The use of ontologies for data integration has been widely studied and has proved its suitability in life sciences (we have also defined and used an ontology for integrating data on genetic variants). Despite its interest, guiding data selection for mining with domain knowledge has not been fully explored. Hence we defined a method involving bio-ontologies for guiding data selection during the preparation step of the KDD process. We defined three scenarios in which SO-Pharm ontology features such as subsumption, properties, class descriptions and domain knowledge, are used for selecting a subset of attributes or objects (patients) for preparing the data mining step. Each of these scenarios was illustrated and evaluated within a case-study relative to the search of genotype-phenotype relationships in a familial hypercholesterolemia dataset [26]. This work will be generalized as an alternative to numerical methods for data selection that evaluate the interest of an attribute as its information gain in the dataset.

6.4.3. Organizing and Querying a Metadata Repository with FCA

The BioRegistry project aims at gathering and organizing metadata about biological data-bases in order to facilitate and to optimize the selection of relevant databases with respect to a user query. Formal Concept Analysis (FCA) was set up for organizing the BioRegistry and visualizing the sharing of metadata across the DB. A formal context representing the relation between bioinformatics data sources and their metadata is provided, and the corresponding concept lattice is built [75]. The BR-explorer algorithm addresses the problem of retrieving the relevant data sources for a given query [17]. Following this work, we studied extensions of the FCA for working on complex multi-valued data (i.e. no longer binary contexts). We first deal with dependencies on attributes in lattice- based Information Retrieval (IR). Introducing dependencies on attributes leads to the definition of hierarchies of attributes in which levels reflect the attribute importance [32]. The distinction between attributes based on their importance is interesting for lattice-based IR both for navigation and for querying. In case of navigation, the choice of moving to a particular concept rather than another may be influenced by the high importance of its attributes. This leads to a navigation guided by domain knowledge. In case of querying, important keywords define the main objective of the retrieval while less important ones define additional information that can be present in the answers. Another ongoing extension of the coupling FCA-IR concerns the management of muti-valued attributes. Actually, metadata fields may have as domain a list of symbolic values belonging to a thesaurus or to a domain ontology (such as the MeSH vocabulary). The challenge here is to build a concept lattice from the metadata in their initial format and to take into account the semantic relationships between the different values.

6.4.4. Drug Design

Bernard Maigret and his students have joined the Orpailleur team on the begining of 2007. Bernard Maigret has been working during the last months within several pending collaborations with different groups [5], [6], [7], [8], [10], [14], [13], [19].

The goal of any drug discovery effort is to develop a chemical product that binds to a target macromolecule known to play a key role in a disease state. Recent progress in genomics, proteomics, high-throughput screening, combinatorial chemistry, molecular biology, and three-dimensional bio-structures determination, has radically changed the approach to drug discovery. At present, typically several millions molecules have to be tested within a short period and, therefore, highly effective screening methods are necessary for lead identification. This situation has created a need for "virtual high-throughput screening" (vHTS) by which researchers can, in silico, prioritize compounds in order to focus resources on those candidates most likely to be biologically active. Our efforts intend to develop the VSM-G platform (Virtual Screening Manager for computational Grids) in order to perform efficient vHTS by integrating a suite of programs to efficiently screen huge molecule databases against protein targets for lead generation and subsequent experimental testing [4]. The core computational technology within VSM-G is a funnel of several receptor-ligand docking methods ranging from fast but crude selection algorithms to must accurate but slower algorithms. All the involved algorithms are easily adaptable on computational grid networks. Our platform should, therefore, act as an efficient filter able to efficiently and rapidly handle millions of molecules and to select the best candidates for another round of more elaborate affinity measurements.

The objective of an ongoing research work consists in discovering docking constraints by mining various annotation data on ligands, targets and known target-ligand bindings. The extracted constraints will be useful either for upstream VSM-G (for pre-filtering the ligand set) or for downstream VSM-G (for ranking the identified candidates). The availability of large-scale prior knowledge on biological entities (proteins, small molecules, molecular complexes) motivates this alternative to vTHS. Moreover the data to be mined are complex and therefore difficult to represent with binary tables that are required for FCA-based methods. Our proposal is to explore relational data mining techniques such as Inductive Logic Programming (ILP) or RCA for achieving the rule extraction process and yielding the docking constraints.

6.5. KDDK and Spatial Reasoning

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

Participants: Charu Asthana, Florence Le Ber, Jean-François Mari, Amedeo Napoli.

In this framework, we work on two major themes, the representation of spatial structures in knowledge-based systems, and the design of reasoning models on these structures e.g. hierarchical classification and CBR. This research work is applied to answer agronomic questions regarding the recognition and the analysis of farmland spatial structures. Besides, we have been involved in the organization of the workshop RTE 2007 on spatial and temporal reasoning (http://afia2007.imag.fr/rte/) [1], and in the co-editing of a book on spatial and temporal reasoning [2].

6.5.1. Lattice-based Classification of Spatial Relations

This work has been initiated during the thesis of Ludmila Mangelinck (1995–1998), in collaboration with the INRA BIA laboratory in Nancy. It has been carried out in the context of the design of a knowledge-based system for agricultural landscape analysis [11].

In this framework, we have designed a hierarchical representation of topological relations based on a Galois lattice –or concept lattice structure– relying on the Galois lattice theory. A Galois lattice is a multi-faceted tool for designing hierarchies of concepts: it allows the construction of a hierarchical structure both for representing knowledge and for reasoning. In a concept lattice structure, a concept may be defined by an *extension*, i.e. the set of individuals being instances of the concept, and by an *intension*, i.e. the set of properties shared by all individuals. In our framework, the extension of concepts corresponds to topological relations between regions of an image, and the intension of concepts corresponds to properties computed on that image regions (*computational operations*). Thus, a concept lattice structure emphasizes the correspondence between qualitative models, e.g. topological relations, and quantitative data, e.g. vector or raster data [12].

Currently, this work is continuing with a deeper study of Galois lattices for linking qualitative topological relations, and computational operations on numerical (raster or vector) data. In particular, we focus on the comparison of lattices built on different sets of relations, or computational operations [18]. In the same way, it can be noticed that concept lattices have also being used for mining and understanding hydrobiological data [23].

6.5.2. CBR on Spatial Organization Graphs

This work has been undertaken in the framework of Jean-Luc Metzger thesis (2000 – 2005), in collaboration with INRA SAD. The objective was to develop a knowledge-based system, called ROSA, for comparing and analyzing farm spatial structures. The reasoning in the ROSA system follows the principles of case-based reasoning (CBR). In our research work, CBR relies on the agronomic assumption that there exists a strong relation between the spatial and the functional organizations of farms, and thus, that similar spatial organizations correspond to similar functional organizations. According to this assumption, and given a set of previously studied farm cases, the ROSA system has to help agronomists to analyze new problems holding on land use and land management in farms. This part of the project is stopped since J.-L. Metzger left the team (end of 2005). Besides, the analysis of the knowledge acquisition and modeling processes, undertaken with the help of researchers in socio-psychology and linguistics (CODISANT, LABPSYLOR, Université Nancy 2 and ICAR UMR 5191 CNRS, Lyon) is continuing [27].

6.5.3. 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. We focused on the experience of people in charge of building "*observatoires*", i.e. information systems for the monitoring and the management of rural territories. Our goal is to build a system based on past experiences to help these people. We rely on previous work on experience-based reasoning [72], story-telling [45] and computer aided design [56].

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 [36], [37], [35]. A database and a Web interface have been developed to allow the consultation of these design episodes. The next step is to record and analyze the way people use the interface.

6.6. Mining with HMMs: Results and Applications

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

Participants: Charu Asthana, Florence Le Ber, Jean-François Mari.

Several applications have been carried out during this year, and one new ANR project in which the Orpailleur team is involved, has been funded. In this project, called "BiodivAgrim" and holding on the domain of "Biodiversité", we are associated with the UR 55 of Inra Sad (Mirecourt). This ANR project has just 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") that will be completed in December 2007.

At the same time, the research project called FOUDANGA within the ACI IMPBIO (for "Informatique, Mathématiques, Physique en Biologie Moléculaire") will be completed. The ADD-COPT project for "Agriculture et Développement Durable", is running in its third year. All these research works have 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. Applications in Agronomy

The goal of the ADD-COPT project – for "Agriculture et Développement Durable"– is to specify an observatory of agricultural practices for supporting the different actors in the transformation process to this new agriculture: allowing these actors to confront and share their knowledge, to apprehend and analyze the observations made on the territory, and to assess the impacts of the changes in progress. In order to have a better understanding of the changes in land use occupations, we have carried out a data mining study on the evolution of the spatial neighborhood between the land use occupations. We showed that the agronomic landscape viewed as a space of land use occupations is isotropic (as far as the region has a surface less than 400 km^2), and that the modeling with a second-order Markov chain of the agronomic piece neighborhood as a function of time gives valuable information to an agronomist expert. For example, we have observed that the farmers tend to keep meadows beside a forest rather a corn field.

A second research project, called BiodivAgrim, is the successor of the ACI ECOGER. It is still lying in the context of the mining of environmental data. It groups together various competences such as agronomy, zoology, and data mining. We are currently using the CAROTTAGE system to process simultaneously time temporal and space data, for allowing the agronomists to analyze data collected during several years on the ground occupation on a set of points in France. Previously, the CAROTTAGE system was already used for understanding the risks faced by bustards, as the disappearance of the meadows clearly impact on their migration. In the ECOGER project, the CAROTTAGE system was used within a broader framework: environmental risks. The software has been adapted to take into account the spatial organization of the successions of land use occupations augmented with topographic and agronomic data such as: pedology, elevation and accessibility, and type of the drainage system of the agricultural pieces under investigation. On these composite data, the stochastic modeling based on HMM2 allowed the identification of homogeneous regions in which the composite data exhibit functional dependencies between their attributes represented by frequent itemsets. A new challenge arises: whereas the software works with numerical temporal and spatial data, it has now to integrate symbolic methods to explore the relationships between the frequent itemsets.

6.6.2. 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 of the “Université Henri Poincaré Nancy 1” in two different areas: the detection of promoters (regulation motifs) and the horizontal transfer understanding. The objective of our work aims at developing a new stochastic approach of modeling the segmentation of the genomic sequences for the identification of heterogeneity in the sequences of ADN. The heterogeneity can arise locally (some nucleotides) or extend on wide areas of several thousands nucleotides. Discrete heterogeneity can correspond to gene regulation motifs, the transcriptional factor binding sites (TFBS) and wider heterogeneity are interpreted as exogenic sequences acquired by mechanisms of horizontal transfer.

In the framework of the detection of promoters, the biological material is the soil-dwelling, filamentous bacteria belonging to the genus *Streptomyces*, that is the greatest source of antibiotics amongst microorganisms. In particular, the 8,7M bases of the *Streptomyces coelicolor* chromosome have been entirely sequenced and annotated. We have detected genome heterogeneity islands, and inter sequences dependencies, using a HMM2 without prior knowledge. This heterogeneity is detected by a local maximum of the posterior probability computed by an adequate HMM2. An enrichment of these motifs in the intergenic areas was observed and suggested that they could correspond to TFBS. We carried out a validation on a subset, consisted of 30 genes known to be controlled by the gene SigR. The TFBS of SigR is composed by two boxes located -35 and -10 nucleotides upstream the regulated gene. All the -35 boxes (GGAAT) were identified whereas some boxes -10 (GTT) remain undetected. In order to achieve a knowledge extraction process (i.e. extracting new TFBS or building sets of co-regulated genes) we have designed a new methodology by means of classification and combinatorial algorithms that analyze the nucleotide areas around the extracted motifs in order to retrieve the most frequent or infrequent short sequences and determine the two boxes that could be the TFBSs. This work has been submitted for publication.

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 *S. thermophilus*. A confirmation or not of the exogenic character of the atypical areas has been undertaken. In addition, the analysis of the genome of CNRZ1066 with Markov models of various orders is under development in order to compare the effectiveness of these models compared to the HMM2 and to determine thereafter, the optimum model describing best the non typicality of the genomes of the lactic bacteria. Lastly, an experimental validation using DNA microarrays could complete the validation of the model.

7. Other Grants and Activities

7.1. The European Network of Excellence Knowledge Web

"Knowledge Web" is the name of a European network of excellence initiated in 2004. Three INRIA teams are involved in Knowledge Web, namely ACACIA at INRIA-SOPHIA, EXMO at INRIA-RHÔNE-ALPES and Orpailleur. The current World Wide Web (WWW) is the syntactic Web, where the structure of the content of documents is presented, while the content of documents itself is inaccessible to computers. The next generation of the Web, the Semantic Web, aims at alleviating such problem, and provide specific solutions targeted to concrete problems. The Web resources will be much easier and more readily accessible by both human and computers, with an additional semantic information in a machine-understandable and machine-processable form. The Semantic Web will have much higher impact on eWork and eCommerce than the current version of the Web already had. Still, there is a long way to go transferring the Semantic Web from an academic adventure into a technology provided by software industry. Supporting this transition process of Ontology technology from Academia to Industry is the main and major goal of the "Knowledge Web" project. This main goal naturally translates into three main objectives, given the nature of such a transformation:

- Industry requires immediate support in taking up this complex and new technology. Languages and interfaces need to be standardized to reduce the effort and provide scalability to solutions. Methods and use-cases need to be provided to convince and to provide guidelines for how to work with this technology.
- Important support to industry is provided by developing high-class education in the area of Semantic Web, Web services, and Ontologies.
- Research on Ontologies and the Semantic Web has not yet reached its goals. New areas such as the combination of Semantic Web with Web services realizing intelligent Web services require serious new research efforts.

More briefly, it is the mission of Knowledge Web to strengthen the European software industry in one of the most important areas of current computer technology: Semantic Web enabling eWork and eCommerce. Naturally, this includes education and research efforts to ensure the durability of impact and support of industry.

7.2. The Eureka GenNet Project

The research and development GenNet project is a European EUREKA-labeled project, involving two industrial societies, namely the French *KIKA medical* society, and the Belgian *Phenosystems* society. Two members of the Orpailleur group are leading a thesis on the integration of clinical and genetic data for mining and pharmacogenomics knowledge extraction. This research work is in progress, and more developments are needed before substantial results may be obtained.

7.3. National initiatives

7.3.1. ACI IMPBIO: the FouDAnGA Project

The FouDAnGA proposal, for "Fouille de données pour l'annotation de génomes d'actinomycètes" has been selected in June 2004 as an ACI IMPBIO project in bioinformatics. This project involves two research teams from LORIA (namely ADAGE and Orpailleur), and the Laboratory of Genetics and Microbiology of the University UHP Nancy 1. Since a number of years, these three teams have been collaborating within the PRST "Intelligence logicielle – Bioinformatique et applications à la génomique" (see hereafter). Being selected as an ACI IMPBIO project has reinforced and structured the initial project, allowing two students to complete their thesis.

The scientific motivation of this project is to extract subsequences from DNA with informative and significant values in molecular genetics. In particular, the signals implied in the gene regulation are under investigation. The models used correspond to the bacteria of the group of the actinomycetes –in particular to Streptomyces– that is the main producer of antibiotics and of metabolites with therapeutic interest, and with Mycobacteries –for example *M. tuberculosis*– that is responsible for tuberculosis.

A steady homogeneous second-order hidden state chain describes discrete heterogeneities distributed with a strong bias in the intergenic regions. The a posteriori observation of the hidden states specifies short DNA loci (5 to 12 pb) corresponding mostly to targets for DNA binding proteins, including transcriptional regulators. The analysis of the Streptomyces coelicolor genome allows the detection of the exact location of all 30 SigR promoters, as well as 92 other known or putative relevant regulatory sequences described so far. These DNA motifs represent about 7,8% of the 3000 extracted from a database corresponding to 1,15 Mb of chromosomal DNA.

7.3.2. ACI IMPBIO: the ISIBIO working group

ISIBio stands for “Information Systems Integration in Biology” is a research project, supported since July 2004 by the Ministry of Research in the framework of the ACI IMPBIO initiative. In this interdisciplinary project, the interest is on the exploration of the role of metadata and ontologies in the integration of information systems in biology. The ISIBIO project reinforces the existing collaborations between people from different disciplines, and stimulate new interactions at both the national and the international levels, by organizing twice a year an international seminar.

The last ISIBio seminar has been held in Nancy (LORIA) on December 3, 2007. For this occasion, Marie-Dominique Devignes and Malika Smail-Tabbone have co-organized an international workshop that was held in conjunction with the 8th WISE international conference in Nancy on December 4-6, 2007. This workshop was entitled "Approaches and Architectures for Web Data Integration and Mining in Life Sciences (WebDIM4LS)"
.

7.3.3. 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 conducted us to get a PHD grant from the INCa institute (Institut National du Cancer) via the canceropole Grand-Est. The PhD thesis will be co-directed by Orpailleur members (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.

7.3.4. Research Projects in Chemistry

- “INCa”: C-Met (2006-2008)
- “ANR cardiovasulaire”: Apeline (2006-2008)
- “ANR blanche”: FAK (2006-2008)
- International collaborations with:

  Brazil (CNRS/CNPq, 2006-2007),

  Spain (CNRS/CSIC, 2006-2007),

  and Algeria (CMEP, Tassili program 2006-2009).

7.3.5. 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.

7.4. The CPER MISN Programme

7.4.1. Theme MBI within CPER MISN

The current CPER MISN includes a transversal topic on "Modeling the Bio-molecules and their Interactions" which is coordinated by M.-D. Devignes (http://bioinfo.loria.fr). The general goal of this starting project is to study how prior knowledge can be taken into account for improving modeling of biomolecules and their interactions and how this can in turn help in modeling biological systems. Four projects involving collaborations with biology or chemistry laboratories kicked off on spring 2007. Two supplementary projects are under reviewing.

7.4.2. Theme TALC within CPER MISN

The research on KASIMIR has benefited of a grant from the CPER (*Contrat Plan État Région*). More precisely, an *operation* has been proposed and accepted in the TALC project, where TALC stands for “Language and Knowledge Processing” (*Traitement Automatique des Langues et des Connaissances*).

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

Books and Monographs

[1] M. BOUZID, F. LE BER, G. LIGOZAT, O. PAPINI (editors). *Actes du 3ème atelier Représentation et raisonnement sur le temps et l’espace (RTE 2007), Plateforme AFIA, Grenoble*, 72 pages, AFIA, 2007.

[2] F. LE BER, G. LIGOZAT, O. PAPINI (editors). *Raisonnements sur l’espace et le temps : des modèles aux applications*, Traité IGAT - Géomatique, Lavoisier, Paris, 2007.

Articles in refereed journals and book chapters

[3] 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$^{\rm o}$ 3–4, 2007, p. 469–487.

[4] A. BEAUTRAIT, V. LEROUX, M. CHAVENT, L. GHEMTIO, M.-D. DEVIGNES, M. SMAÏL-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. (to appear), 2007, http://hal.inria.fr/inria-00186993/en/.

[5] C. BONNON, C. BEL, L. GOUTEBROZE, B. MAIGRET, J. GIRAULT, C. FAIVRE-SARRAILH. *PGY Repeats and N-Glycans Govern the Trafficking of Paranodin and Its Selective Association with Contactin and Neurofascin-155*, in "Molecular Biology of the Cell", vol. 18, 2007, p. 229-241.

[6] N. FLOQUET, S. MOUILLERON, R. DAHER, B. MAIGRET, B. BADET, M. BADET-DENISOT. *Ammonia channeling in bacterial glucosamine-6-phosphate synthase (Glms): molecular dynamics simulations and kinetic studies of protein mutants.*, in "FEBS Letters / FEBS-Letters; FEBS Microbiol Lett", vol. 581, 2007, p. 2981-2987.

[7] N. FLOQUET, C. RICHEZ, P. DURAND, B. MAIGRET, B. BADET, M. BADET-DENISOT. *Discovering new inhibitors of bacterial glucosamine-6P synthase (GlmS) by docking simulations.*, in "Bioorganic & Medicinal Chemistry Letters / Bioorganic and Medicinal Chemistry Letters", vol. 17, 2007, p. 1966-1970.

[8] M. FOUCAUD, E. ARCHER-LAHLOU, E. MARCO, I. TIKHONOVA, B. MAIGRET, C. ESCRIEUT, I. LANGER, D. FOURMY. *Insights into the binding and activation sites of the receptors for cholecystokinin and gastrin.*, in "Regulatory Peptides", vol. 25, 2007.

[9] M. HUCHARD, A. NAPOLI, M. ROUANE-HACENE, P. VALTCHEV. *Mining Description Logics Concepts With Relational Concept Analysis*, in "Selected Contributions in Data Analysis and Classification", P. BRITO, P. BERTRAND, G. CUCUMEL, F. D. CARVALHO (editors), Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin, 2007, p. 259–270.

[10] X. ITURRIOZ, S. E. MESSARI, N. D. MOTA, C. FASSOT, R. ALVEAR-PEREZ, B. MAIGRET, C. LLORENS-CORTES. *Functional dissociation between apelin receptor signaling and endocytosis: implications for the effects of apelin on arterial blood pressure*, in "Archives des maladies du coeur et des vaisseaux", 2007.

[11] F. LE BER. *Reconnaissance de paysages agricoles à l'aide de treillis de relations*, in "Raisonnements sur l'espace et le temps : des modèles aux applications, Paris", F. LE BER, G. LIGOZAT, O. PAPINI (editors), Traité IGAT - Géomatique, chap. 11, Lavoisier, 2007, p. 291–304.

[12] F. LE BER, A. NAPOLI. *Treillis pour le raisonnement spatial*, in "Raisonnements sur l'espace et le temps : des modèles aux applications, Paris", F. LE BER, G. LIGOZAT, O. PAPINI (editors), Traité IGAT - Géomatique, chap. 8, Lavoisier, 2007, p. 225–248.

[13] V. LEROUX, N. GRESH, W. LIU, C. GARBAY, B. MAIGRET. *Role of water molecules for binding inhibitors in the SH2 domain of Grb2: a molecular dynamics study*, in "Journal of Molecular Structure THEOCHEM", vol. 806, 2007, p. 51-66.

[14] V. LEROUX, B. MAIGRET. *Should structure-based virtual screening techniques be used more extensively in modern drug discovery ?*, in "Computers and Applied Chemistry", vol. 27, 2007, p. 1-10.

[15] J.-F. MARI, C. LARGOUET. *Modèles graphiques pour le raisonnement temporel et spatial*, F. LE BER, G. LIGOZAT, O. PAPINI (editors), Traité IGAT - Géomatique, chap. 9, Lavoisier, Paris, 2007, p. 249–271.

[16] J.-F. MARI. *Application des modèles de Markov cachés à la fouille de données spatio-temporelles*, F. LE BER, G. LIGOZAT, O. PAPINI (editors), Traité IGAT - Géomatique, chap. 12, Lavoisier, Paris, 2007, p. 305–316.

[17] 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", 2007, http://hal.inria.fr/inria-00187116/en/.

[18] 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.

[19] I. TIKHONOVA, E. MARCO, E. LAHLOU-ARCHER, I. LANGER, M. FOUCAUD, B. MAIGRET, D. FOURMY. *Validated ligand binding sites in CCK receptors. next step: computer-aided design of novel CCK ligands.*, in "Current Topics in Medicinal Chemistry", vol. 7, 2007, p. 1243-1247.

Publications in Conferences and Workshops

[20] F. BADRA, J. LIEBER. *Extraction de connaissances d'adaptation par l'analyse de la base de cas*, in "Extraction et gestion des connaissances (EGC'2007), Actes des septièmes journées Extraction et Gestion des Connaissances, Namur, Belgique, 23-26 janvier 2007, 2 Volumes", Revue des Nouvelles Technologies de l'Information, 2007, p. 751–760, http://www.loria.fr/~badra/EGC07.pdf.

[21] F. BADRA, J. LIEBER. *Une approche pour représenter les variations entre cas — Vers une application à l'extraction de connaissances d'adaptation*, in "Actes du quinzième atelier raisonnement à partir de cas, RàPC'07, Grenoble", A. CORDIER, B. FUCHS (editors), Plateforme AFIA, 2007, p. 47–56.

[22] R. BENDAOUD, M. ROUANE-HACENE, Y. TOUSSAINT, B. DELECROIX, A. NAPOLI. *Text-based ontology construction using relational concept analysis*, in "Proceedings of the International Workshop on Ontology Dynamics, Innsbruck (Austria)", G. FLOURIS, M. D'AQUIN (editors), 2007, p. 55–68.

[23] A. BERTAUX, A. BRAUD, F. LE BER. *Mining Complex Hydrobiological Data with Galois Lattices*, in "Proceedings of the 18th International Conference on Database and Expert Systems Application - International Workshop on Advances in Conceptual Knowledge Engineering (ACKE'07), Regensburg", 2007.

[24] A. CORDIER, B. FUCHS, J. LIEBER, A. MILLE. *Acquisition de connaissances du domaine d'un système de* RÀPC *: une approche fondée sur l'analyse interactive des échecs d'adaptation — le système* FRAKAS, in "Actes du quinzième atelier raisonnement à partir de cas, RàPC'07, Grenoble", A. CORDIER, B. FUCHS (editors), Plateforme AFIA, 2007, p. 57–70.

[25] A. CORDIER, B. FUCHS, J. LIEBER, A. MILLE. *Failure Analysis for Domain Knowledge Acquisition in a Knowledge-Intensive CBR System*, in "Proceedings of the 7th International Conference on Case-Based Reasoning, Belfast", Lecture Notes in Artificial Intelligence 4626, Springer, 2007, p. 463–477.

[26] A. COULET, M. SMAÏL-TABBONE, P. BENLIAN, A. NAPOLI, M.-D. DEVIGNES. *Ontology-guided Data Preparation for Discovering Genotype-Phenotype Relationships*, in "Network Tools and Applications in Biology: A Semantic Web for Bioinformatics NETTAB 2007, Pisa Italie", 2007, http://hal.inria.fr/inria-00186988/en/.

[27] F. LE BER, S. LARDON, C. BRASSAC. *Suivi et analyse d'un processus collaboratif de modélisation de connaissances spatiales*, in "Actes de la Conférence Québéco-Française de Développement de la Géomatique (CQFD-Géo 2007), Clermont-Ferrand", Actes sur CD, 2007.

[28] 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.

[29] 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.

[30] 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.

[31] M. MAALA, A. DELTEIL, A. NAPOLI. *Distance sémantique entre concepts définis en ALE*, in "Langages et modèles à objets, Toulouse, (LMO'07), Paris", I. BORNE, X. CRÉGUT, S. EBERSOLD, F. MIGEON (editors), Hermès, 2007, p. 117–130.

[32] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Traitement d'attributs inter-dépendants pour la recherche d'information par treillis*, in "Actes des 18èmes Journées Francophones d'Ingénierie des Connaissances, IC 2007 (Plate-forme AFIA 2007)", Cépaudeuès éditions, 2007, p. 109-120, http://hal.inria.fr/inria-00187110/en/.

[33] E. NAUER. *IntoWeb : une plate forme hypertexte d'extraction de connaissances et de recherche d'information*, in "Cinquième colloque VSST (Veille Stratégique Scientifique & Technologique), Marrakech Maroc", 2007, http://hal.inria.fr/inria-00186705/en/.

[34] E. NAUER, Y. TOUSSAINT. *Dynamical modification of context for an iterative and interactive information retrieval process on the web*, in "CLA 2007 - Fifth International Conference on Concept Lattices and Their Applications, Montpellier France", 2007, http://hal.inria.fr/inria-00186704/en/.

[35] S. NOGRY, C. BRASSAC, F. LE BER. *Mettre à jour des épisodes pour contribuer à la conception d'observatoires de pratiques*, in "Résumés des communications - Epique'2007, Nantes", 2007, p. 44–45.

[36] S. NOGRY, F. LE BER, C. BRASSAC. *Modélisation d'épisodes de conception d'observatoires à partir de corpus narratifs*, in "Actes des 18es Journées Francophones d'Ingénierie des Connaissances (IC 2007), Grenoble", F. TRICHET (editor), poster, 2007, p. 327–328.

[37] S. NOGRY, F. LE BER, C. BRASSAC. *Modélisation d'épisodes de conception d'observatoires à partir de corpus narratifs*, in "Actes du 15ème atelier de raisonnement à partir de cas (RàPC2007), Plateforme AFIA, Grenoble", A. CORDIER, B. FUCHS (editors), 2007, p. 103–110.

[38] 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.

[39] M. ROUANE-HACENE, M. HUCHARD, A. NAPOLI, P. VALTCHEV. *Extraction de concepts et de relations en analyse relationelle de concepts (ARC)*, in "XIVièmes rencontres de la Société Francophone de Classification (SFC-07), Paris", O. HUDRY, I. CHARON, G. HÉBRAIL (editors), ENST Paris, 2007, p. 169–173.

[40] L. SZATHMARY, A. NAPOLI, S. O. KUZNETSOV. *ZART: A Multifunctional Itemset Mining Algorithm*, in "Proc. of the 5th Intl. Conf. on Concept Lattices and Their Applications (CLA '07), Montpellier, France", Oct 2007.

[41] L. SZATHMARY, A. NAPOLI, P. VALTCHEV. *Towards Rare Itemset Mining*, in "Proc. of the 19th IEEE Intl. Conf. on Tools with Artificial Intelligence (ICTAI '07), Patras, Greece", Oct 2007.

[42] 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.

References in notes

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

[44] D. FENSEL, J. HENDLER, H. LIEBERMAN, W. WAHLSTER (editors). *Spinning the Semantic Web*, The MIT Press, Cambridge, Massachusetts, 2003.

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

[46] S. STAAB, R. STUDER (editors). *Handbook on Ontologies*, Springer, Berlin, 2004.

[47] A. AAMODT. *Knowledge-Intensive Case-Based Reasoning and Sustained Learning*, in "Proc. of the 9th European Conference on Artificial Intelligence (ECAI'90)", L. C. AIELLO (editor), 1990.

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

[53] B. CALVO, N. LOPEZ-BIGAS, S. J. FURNEY, P. LARRANAGA, J. A. LOZANO. *A partially supervised classification approach to dominant and recessive human disease gene prediction*, in "Comput Methods Programs Biomed", vol. 85, n^{o} 3, 2007, p. 229-37.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

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

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

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

[66] 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.

[67] 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.

[68] F. KAREL, J. KLEMA. *Quantitative association rule mining in genomics using apriori knowledge*, in "Proceedings of the ECML/PKDD07 Workshop Prior Conceptual Knowledge in Machine LEarning and Data Mining", 2007, p. 53-64.

[69] 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.

[70] 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.

[71] 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.

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

[73] 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.

[74] 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.

[75] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Querying a Bioinformatic Data Sources Registry with Concept Lattices*, in "Conceptual Structures: Common Semantics for Sharing Knowledge, Proceedings of the 13th International Conference on Conceptual Structures, ICCS 2005, Kassel, Germany", F. DAU, M.-L. MUGNIER, G. STUMME (editors), Lecture Notes in Computer Science 3596, 2005, p. 323–336.

[76] 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.

[77] 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.

[78] S. ROSSI, D. MASOTTI, C. NARDINI, E. BONORA, G. ROMEO, E. MACII, L. BENINI, S. VOLINIA. *TOM: a web-based integrated approach for identification of candidate disease genes*, in "Nucleic Acids Res", vol. 34, n^{o} Web Server issue, 2006, p. W285-92.

[79] Y. SALIHA. *Recherche de gènes candidats responsables du Syndrome d'Aicardi : complémentarité des approches expérimentales et bioinformatiques*, Thèse d'Université (Spécialité Génétique humaine), Université Henri Poincaré, 2007.

[80] 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.

[81] H. SUN, H. FANG, T. CHEN, R. PERKINS, W. TONG. *GOFFA: Gene Ontology For Functional Analysis - A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data*, in "BMC Bioinformatics", vol. 7 Suppl 2, 2006, S23.

[82] V. SVATEK, J. RAUCH, M. FLEK. *Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality*, in "Proceeding of the ECML/PKDD05 Workshop on Knowledge Discovery and Ontologies 2005", 2005.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.