INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team Orpailleur

Knowledge Extraction

Lorraine

THEME COG

Table of contents

1. Team

Project Home Page : http://www.loria.fr/equipes/orpailleur/

Team Leader

Amedeo Napoli [Directeur de recherche CNRS]

Administrative assistant

Antoinette Courrier [Technicienne CNRS]

Staff members

Marie-Dominique Devignes [Chargée de recherche CNRS]
Florence Le Ber [Professeure, ENGEES Strasbourg]
Jean Lieber [Maître de conférences, Université Henri Poincaré Nancy 1]
Jean-François Mari [Professeur, Université de Nancy 2]
Emmanuel Nauer [Maître de conférences, Université de Metz]
Malika Smaïl [Maître de conférences, Université Henri Poincaré Nancy 1]
Yannick Toussaint [Chargé de recherche INRIA]

Ph.D. Students

Mathieu d'Aquin [doctorant, ATER UHP Nancy 1]
Fadi Badra [doctorant, bourse MERT (October 2005)]
Rokia Bendaoud [doctorante, bourse co-financée Région INRIA]
Adrien Coulet [doctorant CIFRE, Kika médical Nancy]
Sébastien Hergalant [doctorant, ATER Université de Nancy 2]
Nicolas Jay [doctorant, assistant hospitalier (Faculté de Médecine, UHP Nancy 1)]
Sandrine Lafrogne [thèse CNAM]
Mohamed Zied Maala [doctorant, bourse France Télécom Sophia]
Sandy Maumus [doctorante, bourse Région INSERM]
Nizar Messaï [doctorant, bourse Région UHP Nancy 1]
Jean-Luc Metzger [ATER Université de Nancy 2 (until October 2005)]
Frédéric Pennerath [doctorant, enseignant Supélec Metz]
Laszlo Szathmary [doctorant, ATER UHP Nancy 1]
Sylvain Tenier [doctorant CIFRE, INIST Diffusion Nancy]

Visiting scientist

Sergei Kuznetsov [Professeur, VINITI Moscou, Russie, (February, March, and November 2005)]

2. Overall Objectives

2.1. Overall Objectives

The "orpailleur" denotes in French a person who is searching for gold in the rivers. In the present case, gold nuggets correspond to knowledge units and may have two major different origins: explicit knowledge that can be given by domain experts, and implicit knowledge that must be extracted from data sources of different natures, e.g. rough data or textual documents. The main objective of the members of the Orpailleur team is to extract knowledge units from different data sources and to design structures for representing the extracted knowledge units. Knowledge-based systems may then be designed, to be used for problem-solving in a number of application domains such as agronomy, biology, chemistry, medicine, the Web...

The research work of the Orpailleur team may be considered from three main interrelated viewpoints: knowledge extraction, knowledge representation, and semantic Web. First, the data sources are prepared to be processed, then they are mined, and finally, the extracted information units are interpreted for becoming knowledge units. These units are in turn embedded within a representation formalism to be used within a knowledge-based system. The mining processes are based on the *classification* operation, e.g. hidden Markov

models, lattice-based classification, frequent itemset search, and association rule extraction. The mining process may be guided by a domain *ontology*, that is considered as a domain *model*, used for interpretation and reasoning.

The whole transformation process from rough data into knowledge units is based on the underlying idea of *classification*. Classification is a polymorphic process involved in a number of tasks within the data to knowledge transformation, e.g. mining operations, modeling of the domain for designing a domain ontology (or extending the ontology with extracted knowledge units), knowledge representation and reasoning. Finally, the knowledge extraction process and the associated knowledge base can be used for problem-solving activities within the framework of the Semantic Web, e.g. Web mining, intelligent information retrieval, content-based document mining...

2.1.1. Note on the organization of the report.

Regarding the organization of this report, for convenience, applications and scientific results are not presented in specific sections, but, instead, follow the theoretical topics on which they are based.

3. Scientific Foundations

3.1. Knowledge Discovery in Databases

Keywords: *association rule extraction, bioinformatics, data mining methods, frequent itemset search, hidden Markov models for data mining, knowledge discovery in databases, lattice-based classification, text mining.*

Participants: Fadi Badra, Rokia Bendaoud, Adrien Coulet, Marie-Dominique Devignes, Sébastien Hergalant, Nicolas Jay, Florence Le Ber, Jean-François Mari, Sandy Maumus, Nizar Messaï, Amedeo Napoli, Frédéric Pennerath, Malika Smaïl, Laszlo Szathmary, Sylvain Ténier, Yannick Toussaint.

Knowledge discovery is a process for extracting knowledge units from large databases, units that can be interpreted and reused within knowledge-based systems.

3.1.1. Symbolic Methods in Knowledge Discovery

Knowledge discovery in databases (KDD) consists in processing a huge volume of data in order to extract useful and reusable knowledge units from these data. An expert of the data domain, called hereafter the *analyst*, is in charge of guiding the extraction process, on the base of his objectives and of his domain knowledge. The extraction process is based on data mining methods returning information units from the data. The analyst selects and interprets a subset of the units for building "models" that may be further interpreted as knowledge units with a certain plausibility.

The KDD process is performed with a KDD system based on four main components: the databases (or the set of data), a domain ontology (and an associated knowledge-based system), data mining modules (either symbolic or numerical), and interfaces for interactions with the system, e.g. editing and visualization. For handling huge volume of data in a given domain, a KDD system may take advantage of domain knowledge, i.e. an ontology, and the problem-solving capabilities of a knowledge-based system working in the domain of data. In turn, closing the loop, the knowledge units extracted by the KDD system may be integrated within the ontology to be reused by the knowledge-based system for future problem-solving operations.

3.1.1.1. Lattice-based classification, frequent itemset search, and association rule extraction.

Symbolic methods for KDD mainly rely on lattice-based classification, frequent itemsets, and association rule extraction [34]. Lattice-based classification is used for extracting from a database (or a set of rough data) a set of concepts organized within a hierarchy i.e. a partial ordering. Lattice-based classification relies on the analysis of boolean tables relating a set of individuals with a set of properties (or characteristics), where *true* stands for the individual `i` has the property `p`. The lattice may be built according to the so-called *Galois* correspondence, classifying within a formal concept a set of individuals, i.e. the extension of the concept,

sharing a common set of properties, i.e. the intension of the concept. In addition, lattice-based classification is the basic operation underlying the so-called *formal concept analysis*.

In parallel, the extraction of frequent itemsets consists in extracting from boolean tables sets of properties occurring with a support or frequency, i.e. the number of individuals sharing the properties, greater than a given threshold. >From the frequent itemsets, it is possible to generate association rules of the form `A` $\longrightarrow$ `B` relating the subset of properties `A` with the subset of properties `B`, that can be interpreted as follows: the individuals including `A` also include `B` with a certain support and a certain confidence. The number of rules that can be extracted is very large, and there is a need for pruning the sets of extracted rules for interpretation (most of the time, the analyst is in charge of interpreting the results of the rule extraction process). Different measures have been set on, mainly based on probability theory, that can be used for pruning the sets of extracted rules, a kind of "rule mining" (see hereafter the text mining section). Finally, let us mention that the team is currently developing a platform for knowledge extraction, called CORON, that includes collections of data filtering methods, and symbolic data mining algorithms. The CORON [37], [29] platform is used in a number of KDD applications that are described in the following.

3.1.2. KDD in biology and medicine

3.1.2.1. Data integration and knowledge extraction in bioinformatics.

Biological datasets have tremendously grown in size and complexity in the past few years. Genome sequences, biomolecule structures, expression arrays, proteomics represent terabytes of data which are stored under variable formats in dispersed heterogeneous databases (DB). More than 700 such DB have been listed at the beginning of 2005. One of the major challenges in the post genomic era consists in exploiting the vast amounts of biological data stored in those DB. 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, systems biology, pharmacogenomics, etc. The knowledge discovery in biological databases process starts with two important steps: data selection from appropriate DB, and data integration. In the biological domain, these tasks are hampered by at least two distinct problems: (i) identifying the relevant DB, and (ii) managing the complexity and heterogeneity of biological data for their integration. Previous and present work within the Orpailleur group has been dealing with the first two aspects of the KDD process: selection of biological databases and heterogeneous biological data integration.

3.1.2.2. Heterogeneous biological data integration: a generic solution to handle user-designed bioinformatic workflows (The Xcollect project).

Heterogeneous data integration has been addressed from a pragmatic and a user-oriented point of view, that has given birth to the Xcollect software. This software provides a generic solution for automated collecting and integration of biological data given a user-defined workflow. We have also investigated the problem of "homologous" answers retrieved from several DB, e.g. different functional annotations for a given gene retrieved in different data sources, and how the variation in quality between DB, e.g. update frequency, manual revision, may be taken into account when presenting the answers to the user. Xcollect [14] has been instantiated to several biological problems: finding mapping data for given genes, identifying possible PPAR-regulated genes, locating transcription start sites on the genome, finding candidate genes for rare diseases.

3.1.2.3. Selection of biological databases: A metadata repository to organize knowledge about biological databases (The BioRegistry project).

The existing solutions and our experience with the Xcollect system have raised up a special need for gathering and for organizing knowledge about biological databases in order to facilitate and to optimize the selection of relevant databases with respect to a user query. Observed limits are dual: either sophisticated models, e.g. mediation architectures such as TAMBIS hamper large scale instantiation, and thus poorly reflect the diversity of biological databases. More simple models, e.g. portals or catalogs such as DBCAT, represent a wide range of databases but offer limited possibilities of extension. A good example of the needs is given by the web service registries e.g. UDDI, and in bioinformatics, MyGrid or BioMoby that allow web services

discovery on the basis of their description. Not all biological databases yet offer access through web services. However, this situation may change, and such web services discovery systems will later benefit from the efforts in modeling and organizing knowledge about biological databases. We thus decided to build a new registry called "BioRegistry", in which the various metadata attached to biological databases accessible through web services or not, are structured, and whenever possible expressed in terms of domain ontologies.

We have proposed a model for the BioRegistry metadata [28]. According to DCMI ("Dublin Core Metadata Intitiative") recommendations, standard data types are involved wherever possible, for example dates and time ranges at format W3CDTF, and existing controlled vocabularies or domain ontologies are used to valuate metadata fields where appropriate. The *subjects* field for instance contains terms extracted from the biomedical thesaurus MeSH, maintained by NLM. This thesaurus was chosen because it is widely used to index scientific literature, and it presents a broad coverage of many biological domains and is regularly updated to take into account changes in the topics addressed by scientific papers. Concerning the *organisms* field, the NCBI taxonomy of living organisms has been chosen since this taxonomy is also used to annotate biological sequences.

The inclusion of several databases in the BioRegistry repository has been performed manually. To accelerate the process, an automatic procedure was designed to import the DBCAT metadata (DBCAT is a catalog no more maintained). In order to valuate the BioRegistry topic information subsection, several text-mining programs were set up to further exploit the DBCAT content. The constraint here was to translate the DBCAT information into controlled vocabulary terms. Hence, we built a correspondence table between the DBCAT *domain* field and MeSH terms to be included in the subjects subsection of the Bioregistry. Additional MeSH terms indexing the publications referred to in the DBCAT *citation* field were also retrieved from MedLine. Since the DBCAT catalog does not contain any field related to the organisms concerned with the data in a given database, the DBCAT *description* field was parsed to retrieve any matching terms with the NCBI taxonomy. Retrieved terms were entered in the *organisms* subsection of the BioRegistry repository. Automatically created XML files (about 500) are currently being manually checked and curated thanks to an editor, developed as a java application (BioRegistry Metadata Editor), and capable of checking the schema constraints. Additional automatic or semi-automatic procedures to populate and update the BioRegistry will be developed in the next future. Exploitation of the Nucleic Acids Research 2005 catalog of molecular biology databases maintained at NCBI is planned. Alert and survey mechanisms have to be designed to detect any change or new release in existing databases as well as new databases appearing on the Web. Future management of the BioRegistry should be performed by the "INstitut de l'Information Scientifique et Technique" (INIST, http://www.inist.fr). This will consist in hosting and maintaining the repository, and offering a querying interface for the biologists. Ideally in the future, any person involved in the construction or maintenance of a biological database should be able to fill in a BioRegistry submission form on line in order to enter a specific database into the repository.

3.1.2.4. Organizing and Querying the BioRegistry with Concept Lattices.

We rely on the use of Formal Concept Analysis (FCA) for organizing the BioRegistry, and for visualizing the sharing of metadata across the DB [27], [26], [25]. We propose the BR-explorer algorithm, a sound and complete bioinformatic algorithm for biological data sources retrieval, based FCA techniques (mainly lattice-based classification), and domain ontologies. The BR-explorer algorithm addresses the problem of retrieving the relevant data sources for a given query. An exhaustive formal context representing the relation between bioinformatic data sources and their metadata is provided, and the corresponding concept lattice is built. Then BR-explorer algorithm starts by building the query concept representing the query, and then inserting the query concept in the concept lattice. Then, the BR-explorer algorithm fills a list of candidate concepts, i.e. a concept that shares at least a property with the query concept (by exploring the ascendant of the query concept in the concept lattice, until the top concept is reached). Finally, the BR-explorer algorithm returns the set of relevant data sources ranked according to their relevance with relation to the considered query. An ontology-based query refinement procedure is currently under investigation, to be integrated in the BR-explorer algorithm, for taking advantage of the semantics of the queries and of the data sources metadata.

3.1.2.5. KDD in bioinformatics: Knowledge extraction in pharmacogenomics.

Most applications of data mining methods currently concern mining homogeneous biological data like protein sequences or structures. Another challenge is the mining of complex heterogeneous data to discover interactions between genes and environment, or between genetic and phenotypic data. This is the case for both public health domain and pharmacogenomics domain, for which scientists point out the difficulty in integrating such heterogeneous biological data. Ultimately, KDD in bioinformatics should help us in improving our knowledge about systems biology. Hence, another ongoing research work consists in applying the whole KDD process to the pharmacogenomics context, i.e. from data selection and filtering to knowledge extraction guided by the domain knowledge. More precisely, the goal is to discover knowledge about interactions between clinical, genetic and therapeutic data. For example, a given genotype –set of selected gene versions– may explain adverse clinical reactions, e.g. hyperthermy, toxic reaction...to a given therapeutic treatment. The first challenge is integration of both genomic and clinical data into a data warehouse on which knowledge extraction operations have to be applied. This knowledge extraction process on complex data must be guided by domain knowledge, taken from a domain ontology. Moreover, even data selection and filtering has to be guided by domain knowledge. This work, that takes place in the framework of the GenNet European EUREKA-labeled project, shows some convergence with the KASIMIR semantic portal construction (see hereafter): the knowledge base management operations may be reusable, such as knowledge editing, knowledge access, knowledge retrieval... Reciprocally, the knowledge extraction process in the pharmacogenomics project may enrich the portal environment with additional modules for data preparation, data mining –with the CORON platform–, and interpretation of data mining results.

3.1.2.6. Association rule extraction in a biological database.

Relying on the KDD principles, a research work is currently under investigation in the domain of biology for searching associations between biological parameters involving cardiovascular (CV) risk factors in a given population of individuals. The studies carried out here rely on a real-world individual database, the STANISLAS cohort. It is a ten-years study holding supposed healthy French families. Families are examined every five years. At the beginning of the study, in 1993, 1006 families (composed by two parents and at least two children) were recruited for medical examination at the "Centre de Médecine Préventive de Vandœuvre-lès-Nancy (France)". Families have been examined further around 1998–1999, and 2003–2004.

The cohort is explored for searching for genotypes and intermediate phenotypes of cardiovascular diseases (CVD), which are multifactorial pathologies resulting from gene-gene and gene-environment interactions. There is a need for extracting implicit and new potential risk factors for CVD within an always increasing volume of data (mainly due to the development of technologies such as PCR multiplex or microarrays). In the STANISLAS cohort, information holds on environmental, clinical, biological and genetic data. The KDD experiments has given results in accordance with the domain knowledge, and as well, other results allowing new research insights for further investigations [23], [24]. Regarding statistical methods usually used in this context, the general idea of the present research work is to mine the cohort for extracting itemsets that may be in turn considered as hypotheses to be validated by statistical tests.

A number of experiments have been carried out, for extracting knowledge concerning a specific pathology related to CVD, namely the metabolic syndrome (MS), i.e. a group of CV risk factors, that may characterize some people in the STANISLAS cohort. We have used the CORON platform for extracting closed frequent itemsets and association rules. The first results have given some precious information on MS in the STANISLAS cohort. In the next future, a combination of symbolic and statistics data mining methods is planned, for a more efficient and complementary study of the complex data of the STANISLAS cohort.

3.1.2.7. Extracting knowledge in medico-economical databases.

Chronic diseases imply recurrent hospitalizations. In order to optimize healthcare resources, improve cooperation between hospitals treating chronic patients, it is very important to understand the factors that may determine the so-called *pathway* of a chronic patient. The patient pathway may be seen as a time-ordered sequence of events affecting the health of the patient. An event describes a set of informations related to an hospitalization, such as, diagnoses, medical or surgical procedures, hospitalization locations, durations, and

costs... In France, the so-called PMSI (for “Programme de Médicalisation des Systèmes d’Informations”) is the name of the information system collecting for an hospital the informations mentioned above.

At present, we are carrying out a research work on the data collected within the PMSI with the following objectives:

- The discovery of elements that may characterize the patient pathway.
- The classification of patients with respect to their pathway.
- The visualization of the patient pathway.

The first objective relies on the extraction of frequent patterns, sequential and not sequential, from the data of PMSI associated to the Lorraine Region. The database includes informations on more than 800 000 hospitalizations per year. The two following objectives allow, based on the patterns that have been extracted, to build and to visualize a patient pathway classification, using concept lattices (or Galois lattices). More generally, this research work aims at investigating the relations that may exist between frequent itemsets, sequential itemsets, and knowledge representation and visualization with concept lattices. A first experiment has been carried out with data on cancer patients in the town of Nancy [17], [8]. This work has been also used for the medicine thesis of Nicolas Jay (*Analyse de la trajectoire de soins en cancérologie. Fouille de données et Extraction des connaissances à partir du PMSI*, Thèse de docteur en Médecine, Faculté de Médecine, UHP Nancy 1, October 2004).

3.1.2.8. Knowledge discovery in chemical reaction databases.

In this paragraph, we briefly present a research work on knowledge discovery in chemical reaction databases. Chemical reactions are the main elements on which relies synthesis in organic chemistry, and this is why chemical reactions databases are of first importance. >From a problem-solving process point of view, synthesis in organic chemistry must be considered at several levels of abstraction: mainly a strategic level where general synthesis methods are involved, and a tactic level where actual chemical reactions are applied. The research work carried out in the present case is aimed at discovering general synthesis methods from chemical reaction databases in order to design generic and reusable synthesis plans.

A first research work based on frequent levelwise itemset search and association rule extraction, and on chemical knowledge, has been carried on, and has given substantial and promising results. At present, this first research work is extended, trying to adapt a graph-mining process for extraction knowledge from chemical reaction databases, but this time directly form the molecular structures and reactions themselves (both being represented as graphs in reaction databases).

3.1.3. Data Mining with Hidden Markov Models

For designing a complete knowledge discovery system, we have developed stochastic models based on high-order hidden Markov models [11]. These models are capable to map sequences of data into a Markov chain in which the transitions between the states depend on the n previous states according to the order of the model. The following experiments are based on second-order hidden Markov models (HMM2), i.e. the transitions between the states depend on the *two* preceding states, for discovering spatial and temporal dependencies in databases. 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 Markovian model from a corpus of observations and an initial model. The resulting Markovian model is able to segment each sequence of data into stationary and transient parts.

We focused our effort 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. (2) The specification of adapted visualization tools giving a synthetic view of the classification results to the experts who have to interpret the classes and/or specify new experiments.

Several applications have been carried out during this last year, and two ANR projects in which the Orpailleur team is involved have been selected: the ADD-COPT project for “Agriculture et Développement Durable”, and the ECOGER project (for “Écologie pour la Gestion des Écosystemes et de leurs Ressources”). In parallel, the

research project called FOUDANGA within the ACI IMPBIO (for "Informatique, Mathématiques, Physique en Biologie Moléculaire") runs in its second year [33].

All these research works have taken advantage of the CAROTTAGE system, a generic data-mining system for spatio-temporal data, based on HMM2 (the CAROTTAGE system is a free software with a GPL license).

3.1.3.1. Two applications in agronomy.

The first ANR project, called ADD-COPT for "Agriculture et Développement Durable", aims at understanding the agriculture evolution for respecting the environment. An agriculture more respectful of the environment will modify the organization of the territory at several levels, i.e. spatial, economic and organizational levels. In this project, we work in collaboration with agronomists, but also with geographers, since there is a question on the representation of territories, with economists since bioagriculture (organic agriculture) must remain economically viable, and with psychologists, who have to formalize how the different actors may share their knowledge for achieving this common objective of new agriculture. The goal of the ADD-COPT project 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.

A second research project, called ECOGER, 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 at the same time temporal and space data, for allowing the agronomists to analyze data collected during several years on the ground occupation on a whole of points of the French territory. Preceding the ECOGER project, the CAROTTAGE system has been already used for understanding the risks that the bustard was facing while the disappearance of the meadows is clearly impacting its migration. In the ECOGER project, the CAROTTAGE system has to be used within a broader framework: environmental risks. The software has to be adapted to take into account the space organization of the successions of cultures. The challenge is double: whereas the software works with temporal data, it has to integrate spatial dimensions and, whereas it has been already tested on relatively homogeneous data, it has to be able to integrate data at different scales, e.g. satellite images, investigations with farmers...

3.1.3.2. An application in bioinformatics.

In the framework of the so-called *Contrat de plan État-Région* that partially supports the thesis work of Sebastien Hergalant in bioinformatics, we are carrying out a long term data mining project with the laboratory of genetics of the "Université Henri Poincaré Nancy 1". 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. One objective in this research work is to detect genome heterogeneity islands, and inter sequences dependencies, using hidden Markov models without prior knowledge. Initially, the focus was put on the understanding of horizontal transfer phenomena [15], but another area of interest in genetics is currently under investigation, namely the detection of promoters [16].

- The horizontal transfer understanding. Markovian models with "specie specific homogeneity" have been constructed, and coupled with transform filters. Their behavior generates regions with different statistical properties, allowing the user to separate "foreign DNA regions" with the own DNA regions of the studied specie itself. In *Streptomyces coelicolor*, the regions with such a statistical consensus have been detected, and correlated with potential events of horizontal transfer.
- The detection of promoters. This work is mainly part of the forthcoming thesis of Sébastien Hergalant (to be defended during Spring 2006). A data mining method based on second order hidden Markov models (HMM2) that is able to process the whole genome sequences without prior hypotheses has been applied to the actinomycete genomes of Streptomyces and Mycobacterium species. The stochastic modeling of the genome with HMM2 allows the extraction, and the classification of short segments (5 to 12 bp) having a significant different structure and composition without any prior knowledge. These segments appear to be parts of binding sites for transcriptional factors.

- In order to confirm the applicability of this new method to the detection of transcriptional signals, the models have been applied to an experimentally determined co-regulated gene set (30 genes) dependent on SigR, a sigma factor of Streptomyces coelicolor involved in the oxidative stress response. A steady homogeneous second-order hidden state chain describes discrete heterogeneity visualized as peaks in the a posteriori observation of the hidden states.
  The duration capabilities of the HMM2 shows very good performances in the modeling of short segments such as TFBS (Transcriptional factors binding sites), and RBS (Ribosome binding sites). On different genomes of the actinomycete family, the HMM2 reveal DNA heterogeneity, that are combined to predict known or potential TFBS or RBS. Based on these models, the present data mining method has proved to be efficient for the detection of DNA motifs involved in both transcriptional, i.e. sigma factor binding sites, and translational (RBS) regulation.

3.1.4. The text mining process

The objective of a text mining process is to extract new and useful knowledge units in a large set of texts. The text mining process relies on the principles of KDD, although it shows some specific characteristics due to the fact that texts are 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 paraphrases, ambiguities, specialized vocabulary, and terminology. This is why the first steps of a text mining process are usually dedicated to linguistic knowledge acquisition: lexicon, terminology, markers of semantic relations, discourse markers, specific syntactic or semantic structures...The following steps of the text mining process aim at identifying or structuring the background knowledge for extracting new knowledge units. The interpretation of a text relies on a common knowledge shared by the authors and the presumed readers. Part of this background knowledge is expressed in the texts, and, even it can be extracted by the text mining process, cannot be considered as new knowledge. Another part of the background knowledge is not explicitly expressed but is useful to relate notions in a text that, at a first glance, seem to be disconnected.

To carry out studies on text mining, the Orpailleur team is interested in linguistic resources, working on real-world texts in application domains such as biology, astronomy..., using robust tools, contrasting with other mining approaches dealing with specific language phenomena. Here, the language is considered as a way for accessing information, and not as an object to be studied for its own. Thus, the text mining process is involved in a loop, that can be used to improve and update linguistic resources. In turn, linguistic resources can be exploited to improve the knowledge extraction process, and for guiding a “knowledge-based text mining process” or KBTM, where “knowledge” makes reference to the knowledge available on the domain of texts. Moreover, as in a standard KDD process, an analyst is usually in charge of supervising the text mining process.

3.1.4.1. Extraction of association rules from texts.

A number of experiments on the extraction of association rules from texts has been carried out, in the context of scientific and technological watch. One major problem is that the extraction process generates a very large number of rules. Then, selecting an “interesting rule” involving new knowledge units is a rather complex task for an analyst. Thus, the extraction process is considered from two points of view: ranking and evaluation of the rules [3], [32], [12], [6], [5].

- Ranking association rules with statistical indexes.
  Texts are annotated (or indexed) using a thesaurus (or a set of several thesauri when available in the application domain), or a domain ontology. The text mining process relies on an association rule extraction process. The support and the confidence are the two main indexes that are associated to the rules, and they play a major role in the reduction of the calculation time and of the number of the generated rules. These indexes, together with other indexes, e.g. the interest, the dependency, the novelty have been studied. The rules, ranked according to these indexes, have been presented to the analyst. It turns out that some combinations of these indexes to rank the association rules allow the analyst to identify complex semantic relations between terms or synonyms. We have proposed

an algorithm to combine these indexes for classifying the extracted rules. However, most of the extracted rules are in accordance with the present domain knowledge, and do not carry any new knowledge aspect. Thus, another objective is to extract association rules providing new knowledge aspects, using a domain ontology.

- Ranking association rules with a domain ontology.
  The text mining process may be performed with respect to a domain ontology. Text mining may be then considered as an interactive and incremental process, where the domain ontology is used to rank the association rules, and where the association rules are used to enrich and update the existing ontology. We have defined a "likelihood measure" for ranking the association rules, with respect to a "degree of novelty" that a rule may include, compared to the existing background knowledge. This research work is still under development, and has to be extended in a number of directions, e.g. exploiting a complex ontology, defining a likelihood measure adapted to a given domain...
- Structuring association rules within hierarchies.
  In a similar way, another research work aims at extracting knowledge units from texts, for extending a domain ontology. An experience in astronomy is currently under development. Association rules may be, after extraction, organized within a rule hierarchy, to be considered form several points of view. Subsumption on rules may be based on the subsumption relation defined in the Galois or concept lattice of the associated close itemsets underlying the rules (recall that an association rule is extracted from a frequent possibly closed itemset). In parallel, a knowledge model structuring the items within a specialization hierarchy, may also be exploited for guiding the building of the association rule hierarchy.

3.1.4.2. Knowledge extraction from Web pages.

This research work is concerned with the design of a system for extracting knowledge from Web pages that is used as a "semantic annotations" for manipulating the documents by their content [30]. The present approach involves a wrapper-based machine learning algorithm combined with a reasoning process, taking advantage of a domain ontology. The ontology is implemented within the Web Ontology Language (OWL) (reasoning mechanisms such as classification and subsumption are available).

The global context of the present research work is the study of research themes within the European Research Community. The objective is to use the information provided by research teams on their website to generate knowledge about the European Research Community, for technological watch, analysis of research themes, or detection of new research directions.

More precisely, the semantic annotation of an element in a Web page relies on two main operations: (i) identification of the syntactic structure of a specific element in the Web page using a syntactic wrapper-based approach, (ii) identification in an associated domain ontology of the most specific concept subsuming the extracted element, that will be used for building the annotation. Given a document to be analyzed, the system detects with a specific wrapper satisfying given constraints the elements in the document that have to be analyzed and semantically annotated. Then, the detected element is represented as an individual (within the representation language) to be classified as an instance of a concept of the ontology. The classification process may also be used to complete the description of the extracted individual on the basis of additional contextual information that may be given by the wrapper. At the end of the classification process, the element in the document may be represented and annotated using the concept of which the extracted individual is an instance of.

3.2. Knowledge Representation, Knowledge Systems and Semantic Web

Keywords: *case-based reasoning, classification-based reasoning, description logics, knowledge representation, lattice-based classification, nowledge-based information retrieval and extraction, object-based representation systems, qualitative spatial reasoning Semantic Web*.

Participants: Mathieu d'Aquin, Fadi Badra, Florence Le Ber, Jean Lieber, Sandrine Lafrogne, Jean-Luc Metzger, Amedeo Napoli, Emmanuel Nauer, Laszlo Szathmary, Yannick Toussaint.

Knowledge representation is a process for representing knowledge within knowledge a 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. in taking into account the semantics of the elements included in the documents.

3.2.1. Classification-based Systems and Reasoning

A knowledge system relies on a knowledge base and a reasoning module for problem solving and knowledge management in a given domain. Knowledge units are represented within a knowledge representation formalism where they have a syntax and a semantics. Inference can be drawn from already known knowledge units (or facts) for deriving new facts, that are useful for solving the current problem. Moreover, the units extracted from data by data mining procedures have to be represented within a knowledge representation formalism to be taken into account in the framework of a knowledge system.

In the team Orpailleur, two kinds of formalisms are particularly studied, namely description logic (DL) systems, and object-based knowledge representation (OBKR) systems. In these systems, knowledge units are represented within concepts (also called classes), with attributes (properties of concepts, or relations, also called roles in DL), and individuals. The hierarchical organization of concepts relies on a subsumption relation that is a partial ordering. These systems provide inference services such as subsumption, concept and individual classification. Concept classification is used to insert a 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 instance of. In both cases, subsumption and classification are the main operations: this is why these systems are denoted here by "classification-based systems".

Classification-based reasoning may be extended into case-based reasoning (CBR), that relies on three main operations: retrieval, adaptation, and memorization. A source case `(srce,Sol(srce))` lies in a case base, and can be seen as a problem statement `srce` together with its solution `Sol(srce)`. Then, given a new target problem, say `tgt`, retrieval consists in the search for a memorized case whose problem statement `srce` is similar to the target problem `tgt`. Then, when `srce` exists, its solution `Sol(srce)` is adapted to fulfill the constraints attached to `tgt`. When there is enough interest, the new pair `(tgt,Sol(tgt))` can be memorized as a new case for further problem solving. In the context of a concept hierarchy, retrieval and adaptation may be based both on classification. Moreover, a number of studies within the Orpailleur team has been carried out on CBR, especially on "adaptation-guided retrieval", that consists in searching for a source case whose solution will be adaptable for the target problem, giving a kind of guarantee regarding the building of the solution of the source case.

In parallel with knowledge representation, knowledge management is oriented toward the management of what could be called the "cycle" of knowledge, including acquisition, memorization, retrieval, maintenance, dissemination (or exchange) of knowledge. There is also a need for coupling knowledge with data, with respect to representation and management. This means in particular that, besides knowledge extraction from databases, there are some other needs such as e.g. information retrieval, for helping a reasoning process. Thus, there must exist channels between the knowledge representation universes and the document (or data universe). This is particularly important in the framework of the semantic Web introduced hereafter.

These channels can rely on a coupling of a knowledge representation formalism and a description language for documents, such as XML. In this way, knowledge representation units can be associated to document descriptions units: the management of documents (or data) is performed within the document description language, and reasoning is performed within the knowledge representation formalism. Moreover, additional

coupling between information retrieval and knowledge extraction can be set on. This view of knowledge management is of primary importance, mainly because of the Web, and the always growing need of disseminating information and knowledge.

3.2.2. The Semantic Web framework

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.

The semantic Web has gained a great interest in the research work of the Orpailleur team. Indeed, it constitutes a good platform for experimenting a number of ideas on knowledge representation, reasoning, knowledge management, and knowledge discovery (and especially text mining) as well. Investigations mainly hold on the content-based manipulation of textual documents using annotation, ontologies, and a knowledge representation language. The idea is to build an XML-based "bridge" between documents, and the knowledge units of the domain of documents, lying in domain ontology. The annotations attached to documents, and the queries, are built with the help of the concepts of the domain ontology. Then, the manipulation of annotations, e.g. information retrieval, query answering, reasoning on the content of documents, is left to the reasoning module associated to the knowledge representation formalism.

3.2.3. Knowledge Management in Medicine: the Kasimir System

The objective of the KASIMIR research project is 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* or CAV, Vandœuvre-lès-Nancy) and Oncolor (an association of physicians from Lorraine involved in oncology).

For a cancer localization, e.g. the breast, the treatment is based on a protocol similar to a medical guideline. This protocol is built according to evidence-based medicine principles. For most of the cases (about 70%), a straightforward application of the protocol is sufficient, and provides a solution, i.e. a treatment, that can be directly reused.

A case out of the 30% remaining cases is "out of the protocol", meaning that either the protocol does not provide a treatment for this case, or the proposed solution raises difficulties, e.g. contraindication, treatment impossibility, etc. For such an out of the protocol case, oncologists try to *adapt* the protocol (actually they discuss such a case during the so-called "breast therapeutic decision meetings", including experts of all domains 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.

3.2.3.1. 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 three directions: the AKA from experts, the automatic AKA and the 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. These patterns involve new knowledge representation needs. For example, one of these patterns corresponds to the adaptation of protocol, when some crucial pieces of information are missing [13].

Automatic AKA is based on the "mining of the protocols". A protocol can be seen as a set of rules "`situation`$\longrightarrow$`decision`". Knowing how the decisions change when the situations change from one rule to another rule provides a specific adaptation rule. By generalizing these specific rules, general adaptation rules may be obtained. This generalization process has been implemented thanks to a frequent close itemset extraction module of the CORON platform (see § 4.1). 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.

These two kinds of 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). The goal of the semi-automatic AKA will be to combine these two kinds of AKA in order to produce operational *and* intelligible adaptation knowledge units.

The research in AKA has been carried out in the interdisciplinary project TCAN (see section 5.3.4) and will be continued afterward.

3.2.3.2. Knowledge representation for decision support tools.

Two versions of KASIMIR are currently used: one based on an *ad hoc* object-representation formalism (OBRF), the other one based on semantic Web principles, in a semantic portal (as explained below). A number of knowledge bases corresponding to specific cancers (decision protocols) has been developed. Moreover, the inference engine has been extended for taking into account a fuzzy representation of concepts and fuzzy hierarchical classification. The system tries to detect and to propose more than one treatment for "borderline cases": this has been implemented for the OBRF version of KASIMIR, and its implementation in the semantic portal is planned [38]. Another study is about "multiple viewpoint representation and reasoning", that may be useful for modeling the reasoning of the breast therapeutic decision committee, i.e. each viewpoint represents a domain in breast oncology. In [31], the formalism C-OWL for the representation of multiple contextualized ontologies in the semantic Web is adapted for the purpose of multiple viewpoint representation and decentralized CBR.

3.2.3.3. A semantic portal for oncology.

The current research in computer science on the KASIMIR system follows two main directions: protocol adaptation, and the embedding of the KASIMIR system within a semantic portal for oncology, i.e., a Web server relying on the principles and technologies of the semantic Web for providing an intelligent access to knowledge and services in oncology.

One of the main issues of the semantic Web relies on interoperability for knowledge and applications. Thus, building a semantic portal implies a standardization of knowledge and software components of the KASIMIR system. For the knowledge bases, standardization relies on a sharable domain model, and leads to the definition of general ontologies in oncology. This kind of "knowledge base re-engineering" requires to replace the *ad hoc* knowledge representation formalism of KASIMIR with OWL, the knowledge representation formalism of the semantic Web. The representation of protocols is also re-engineered in order to take a better advantage of the expressiveness of the OWL formalism.

This work also implies a new software architecture, for the KASIMIR reasoner and the editing, visualization and maintenance modules [7]. This architecture must take into account constraints related to the distributed and dynamic environment of the semantic Web. In order to query the protocols represented within OWL, an instance editor called EDHIBOU has been developed. Another interface, called NAVHIBOU, has been 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 the 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 forthcoming thesis of Mathieu d'Aquin).

3.2.3.4. Going further: knowledge discovery for the semantic Web.

The semantic portal of KASIMIR is operational in the sense that, given a decision protocol represented in OWL, and an adaptation knowledge base, it can be used to apply or to adapt the protocol to specific situations. Besides, some ongoing research in the KASIMIR project aims at acquiring knowledge, especially adaptation knowledge, as explained above.

This is the goal of the thesis of Fadi Badra, initiated in October 2005, to combine these two research issues, i.e. how knowledge discovery techniques can be used to feed a semantic portal, and how the knowledge server embedded in this portal can be used to assist the knowledge discovery processes.

From a longer term perspective, the goal is the following: having a clear distinction between the notions of data and knowledge, try to build a distributed system, with knowledge bases, heterogeneous data bases, inference engines, knowledge discovery modules, allowing communications with human beings, such as experts and end-users.

3.2.4. Spatial Knowledge Representation and Spatial Reasoning

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 [4]. Besides, we have been involved in the AS 144 ("Action spécifique") of CNRS–STIC [21], and in the organization of the workshop RTE 2005 on spatial and temporal reasoning [1].

3.2.4.1. 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.

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.

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. The representation of relations and structures is currently under investigation [22].

3.2.4.2. CBR on spatial organization graphs.

This work has been undertaken in the framework of Jean-Luc Metzger thesis, in collaboration with INRA SAD and ENGREF. The objective is 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 [9].

- In a first step of the present work, a model of the domain knowledge has been proposed, in accordance with agronomists. This model is based on *spatial organization graphs*, or SOG, with labeled vertexes and edges. Relying on these spatial organization graphs, *spatio-functional cases* for farms have been designed: they mainly consist of a description of the land use, and an associated explanation linking spatial and functional organizations.
- In a second step, the SOGs and the cases have been represented within a knowledge representation formalism, namely the description logic (DL) system RACER. In this way, reasoning in the ROSA system relies on an original combination of hierarchical classification, CBR, and qualitative spatial reasoning. In addition, spatial inference rules are used for building *similarity paths* between SOGs. These paths are used in the CBR mechanism for comparing problems and adapting the solution from a source case to a new target problem.

The knowledge acquisition and modeling issue has been undertaken with the help of researchers in socio-psychology and linguistics (CODISANT, LPI-GRC, Université Nancy 2 and GRIC UMR 5612 CNRS, Lyon) [20].

During this year, the ROSA system has been experimented by the agronomists, and improved. Cases and transformation rules have been formalized and added to the system, while Jean-Luc Metzger has defended his thesis on the first of April 2005 [2].

3.2.5. Intelligent Access to Information

The availability and retrieval of information is of main importance in scientific and technical domains, e.g. for research and technological watch purposes. Nowadays, there is a large quantity of data available, and this requires to implement adapted tools for exploiting this mass of data. A research work holds on the definition and implementation of an environment –a toolbox– allowing an intelligent access to information, by combining information retrieval, hypertext navigation, and data-mining. This toolbox can be used for document retrieval on the Web, bibliographical search or domain analysis.

In this framework, the design of a semantic-based algorithm for comparing and classifying documents is under investigation. The annotations of documents are represented as labelled trees, where nodes and edges are composed of concepts lying in a domain ontology associated with the topics of the considered documents. A reasoning process based on classification is carried out for comparing the labelled trees representing documents, i.e. the annotations, and thus for comparing the documents. This comparison process allows to compute a semantic similarity measure between documents, and then to classify documents according to their content.

Another important idea underlying the toolbox is that data-mining and information retrieval are complementary tasks for accessing and analyzing data. Data-mining allows the guiding of information retrieval by taking advantage of the knowledge units extracted from the data, for example the extraction of a lattice from the data may provide an organization on which the information retrieval process may rely. Conversely, information retrieval allows the guiding of the data-mining process by making available information on data that can be used for example for pruning a set of extracted rules, or for providing a focus for a classification process.

A toolbox, called ”IntoWeb”, is an extension of the so-called “IntoBib” system –a generic system designed for the exploitation of bibliographical data– and provides a set of tools for implementing the core tasks of the knowledge extraction process (see 4.6.1). For building a generic knowledge-based extraction system, it is needed to precisely define the kinds of objects to be exploited, and the operations to be applied to the objects. The objects may be of different types, among others URL (reference to a web document), hypertextual document (web document), full-text document, XML document, and also association rule (correlation between properties), vector (set of valuated properties), ontology...Operations may be applied to these objects for producing new objects, containing some information, e.g. that is searched for, allowing an interrelated and iterative data mining/information retrieval process. Such operations are the retrieval of all the hypertextual documents identified by a set of URL, the computation of the the vectorial representation of a full-text or an XML document, the extraction of an annotation tree from a textual document according an ontology, the

extraction of a set of association rules from a set of XML documents, the classification of web documents according to an ontology [36], [35], etc.

Regarding this framework, this year, we focused on the specification of the IntoWeb system, and have initiated collaboration with the Cortex team in LORIA on information filtering. The mechanism modeling a user profile for content-based filtering systems is based on a "novelty detector" model, that learns the evolutive user needs according to positive and negative user's relevance feedbacks [18], [19]. The next step is now to enhance the effectiveness of this model by taking into account knowledge about the domains in which information has to be retrieved.

4. Software

4.1. A Data Mining Toolkit: the Coron Platform

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

Participant: Laszlo Szathmary [contact person].

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, called *frequent closed itemsets* (FCIs). In turn, FCIs can also be used for finding useful or informative association rules.

We have developed a collection of programs for data mining that are grouped together in the so-called CORON platform. While last year only two algorithms were implemented, namely Close and Titanic, by now more than 10 algorithms are available in the platform. The platform contains well-known algorithms in the data mining community, such as Apriori, Apriori-Close, Pascal, Charm, Eclat, but it also includes several original algorithms such as Pascal$^+$, Zart, RMS Carpathia, Eclat-Z. The toolkit is composed of three main parts: (i) CORON-base, (ii) ASSRULEX, (iii) pre-processing and post-processing modules.

With CORON-base, it is possible to extract different kinds of itemsets, e.g. frequent itemsets, frequent closed itemsets, minimal 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? Among all possible rules we can extract some useful subsets, e.g. informative rules, generic basis, informative basis, etc.

The CORON toolkit supports the whole life-cycle of a data mining task. We have modules for cleaning the input dataset, and reduce 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 for a given viewpoint.

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, in the CABAMAKA project (which is part of the KASIMIR system, see § 3.2.3). Moreover, the CORON implementation of the *Titanic* algorithm has been integrated into the GALICIA 2.0 platform, that is developed at the University of Montreal, Canada. It is also planned to conduct text-mining experiments with CORON in the near future.

4.2. Stochastic systems for knowledge discovery and simulation

Keywords: *Hidden Markov models, stochastic process.*

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

4.2.1. 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 [10].

4.2.2. 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 vegetal 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 transgenes diffusion. 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.

4.3. tamis: A software for text and rule mining

Keywords: *association rule extraction, frequent pattern extraction, knowledge discovery from databases, text mining*.

Participants: Hacène Cherfi, Yannick Toussaint [contact person].

The system, called TAMIS for "Text Analysis by Mining Interesting ruleS" is currently under development. This system allows the navigation through a large set of association rules, such as those produced by a text mining experiment. The TAMIS system is based on a user-friendly interface, and it can be easily used by non-computer scientists, e.g. analysts, experts in the domain of the analyzed data. The association rules are extracted by a mining algorithm, e.g. using the CORON platform in the present case, encoded in a predefined XML format. The TAMIS system stores the rules in a database, and proposes eight different statistical measures for sorting the rules, e.g. support, confidence, interest, conviction, dependence... In this way, the analyst may focus on smaller sets of interesting rules satisfying a given set of constraints. These constraints may be expressed by means of operations on the values of the statistical measures, and on the content of the left/right hand side of a rule.

4.4. Software for Spatial Reasoning

Keywords: *land organization, qualitative spatial reasoning, typological relations*.

Participants: Florence Le Ber [contact person], Jean-Luc Metzger.

Rosa, for "Reasoning on Organization of Space in Agriculture", is a system developed in collaboration with agronomists, whose objective is to record and to maintain an agronomic knowledge base on farms, and to solve problems in agronomy, based on this knowledge base. Two kinds of knowledge elements are considered: domain knowledge, and knowledge on spatial organization and functioning of specific farms. The domain knowledge is described by a hierarchy of spatial concepts and relations (spatial occupation and relations). The spatial organization of farms is described by the so-called "space organization graphs" (SOGs) linking spatial entities through spatial relations. A vertex of a SOG (either a spatial entity or a relation) is labeled and linked to a concept of the domain knowledge hierarchy. The functioning of farms is described within "explanations"

attached to SOGs. An explanation holds on a particular function of the considered farm organization and functioning. The association of a particular SOG with an explanation composes a case, to be used within a case-based reasoning process. The Rosa system is under development, and is implemented within the RACER description logic system.

4.5. The Kasimir System

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

Participants: Mathieu d'Aquin, Fadi Badra, Sandrine Lafrogne, Jean Lieber [contact person], Amedeo Napoli.

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. The ontology editor PROTÉGÉ has been customized for editing the KASIMIR protocols, and it has been connected with the KASIMIR inference engine. The use of the PROTÉGÉ editor involves a simplification of the protocol editing, and the detection of errors during the editing, thanks to the inference engine.

Two visualization modules have been integrated in PROTÉGÉ, allowing the display of the KASIMIR hierarchy of concepts representing the protocol being edited: PALÉTUVIER and HYPERTREE (HYPERTREE has been initially developed in the ECOO team at LORIA [7]). The combined use of these two visualization modules, and of the classical tree widget of PROTÉGÉ, provides several useful features for hierarchy visualization, navigation, and global or focused views.

Finally, a maintenance module has been developed and integrated into PROTÉGÉ, that compares two versions of a protocol in order to separate changed and unchanged elements. This module can be used in particular during an editing session, to visualize the modifications since the beginning of the session.

Actually, two versions of KASIMIR are currently used: one version is based on an *ad hoc* object-based representation formalism, and the other version is developed within the semantic portal, as introduced in the section 3.2.3. This latter is based on OWL and on some extensions of OWL, and has motivated the development of the two user interfaces, namely EDHIBOU and NAVHIBOU, presented above. The software CABAMAKA (see also section 3.2.3) for case base mining for adaptation knowledge acquisition is part of the KASIMIR system.

4.6. Softwares for the manipulation of documents for the Semantic Web

Keywords: *Web crawling*, *information access*, *information retrieval*, *semantic Web*.

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

4.6.1. Intelligent Access to Information

Two systems are under development. A first system, called "IntoBib", is a generic system designed for the exploitation of bibliographical data. Two kinds of objects are manipulated within the IntoBib system, namely bibliographical references and properties –or points of view– about these references, e.g. authors, keywords...The available operations on these specific objects are references filtering using one or more points of view, conceptual clustering of similar references with respect to a given point of view, and extraction of correlation between references. Accordingly, the IntoBib system is based on a toolbox providing a number of modules, among which, hypertext navigation, retrieval of bibliographical references, extraction of correlation between references, search for equivalent references (duplicates), conceptual clustering of similar references, normalization of fields e.g. author name, keywords...

The second system, called "IntoWeb", extends the IntoBib system. The objective is to provide a more generic environment for an intelligent access to information, by combining information retrieval, hypertext navigation, and data-mining. The IntoWeb system contains a set of tools implementing the core tasks of

a knowledge extraction process, i.e. collecting, filtering, and mining data. Solving a given problem of information retrieval, or data mining, is performed by a well chosen sequence of operations that are available in the system.

4.6.2. DefineCrawler: a Generic Crawler

The "DefineCrawler" system can be seen as an information retrieval "meta-system", in the sense that it can be parameterized for satisfying different information retrieval tasks. The DefineCrawler system is based, on a classical information retrieval architecture, and on search engines available on the Web. A number of parameters have been retained, to be adjusted within an XML file for implementing and controlling different information retrieval system behaviors.

- Initialization parameters (`Start`) include the maximum depth of the crawl (`Depth`), a set of starting points for navigation (`URL`, possibly making reference to the URL of a search engine), the directory where have to be stored the data collected by the crawler (`Directory`), the number of parallel processes crawling the Web (`NbThread`), a halting condition (`Stop`) making possible the specification of a maximal crawling time, and thus ensuring a termination of the information retrieval process.
- Validation parameters (`Validation`) include a set of conditions (connected by boolean operators) that must be satisfied by the documents, for eliminating documents without interest with respect to the query, e.g. documents that do not satisfy some criteria, that are not in a fixed language...
- Evaluation parameters within which additional conditions can be set, in order to evaluate the returned 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.

Every validation and evaluation condition 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...

5. Other Grants and Activities

5.1. 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-processible 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.

5.2. 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 drive a so-called "thèse Cifre" on the integration of clinical and genetic data for mining and pharmacogenomic knowledge extraction. This research work is in progress, and more developments are needed before substantial results may be obtained.

5.3. National initiatives

5.3.1. 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 CPER "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.

5.3.2. aci impbio: the isibio project

The ISIBIO project 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 first seminar has been organized in Paris on March 29–30th 2005 (http://bioinfo.loria.fr/Members/devignes/ISIBIO/ISIBio-Programme-29-30mars2005). The ISIBIO project has also sponsored the workshop "Ontologie, Grille et Intégration Sémantique pour la

Biologie", held during JOBIM conference in Lyon, on July 4th, 2005. A second seminar is planned for the end of 2005.

5.3.3. aci "Masse de données en Astronomie"

This research project "Knowledge Discovery and Ontology Design in Astronomy" is carried out in collaboration with the CDS in Strasbourg ("Centre de données astronomiques de Strasbourg"), and the IRIT computer science laboratory in Toulouse. Researchers in astronomy use every day an information network made of journal articles available under an electronic form, and a number of databases, such as the SIMBAD database recording bibliographical entries and measure sets on about three millions of astronomical objects, and the catalog server VizieR recording astronomical catalogs and measure tables published in the astronomical journals. Interested researchers should have access to the content of documents, e.g. journal articles, astronomical object catalogs, or measure tables. For facilitating this access, researchers in astronomy have at their disposal a base of the so-called UCD for "Unified Content Descriptors", i.e. a hierarchical database that has been extracted and designed at the CDS from the content of astronomical catalogs and tables.

The research work currently carried out in collaboration with the CDS concerns the study and the design of an ontology for representing the UCD and astronomical objects as well, starting from a collection of articles –thus involving text mining– and for extending the UCD base. This ontology will be used for a number of important and different tasks for researchers in astronomy, such as intelligent information retrieval based on the content of documents, information manipulation for matching and comparing the content of the astronomical documents. This research work can be seen as a contribution to the research works on the Semantic Web, where the purpose is to attach semantics to astronomical documents, for defining an annotation method of astronomical documents, and for a knowledge-based information retrieval method in heterogeneous astronomical sources.

This year, a methodology for building an OWL ontology of the UCDs has been proposed [36], [35]. The specific task in which this ontology has been used is for retrieving the UCDs representing at the best the description of an astronomical object given by a set of properties. An approximate 2-step classification process is performed by exploiting the metadata linking lexical items used in the descriptions of astronomical objects, and concept properties defining the UCDs in the ontology. The recognition of composed UCDs depending on several concepts has to be studied further. The classification of simple and composed UCDs presents similarities to the works on disjunctive classification, where concepts are defined by union of properties: in this case, owning a subset of properties for an object is sufficient to be classified as an instance of the concept.

5.3.4. cnrs tcan Project

A research work on Adaptation Knowledge Acquisition (AKA) for the KASIMIR system (see section 3.2.3.1) is carried out in the framework of the CNRS interdisciplinary project TCAN for "Traitement des connaissances, apprentissage et NTIC". The objective of AKA is to provide knowledge in the form of *adaptation meta-rules*:

- Automated AKA is based on the mining of the protocols. A protocol can be seen as a set of rules `situation` $\longrightarrow$ `decision`. Knowing how the decisions change when the situations change from one rule to another rule may provide a specific adaptation rule. Clustering and generalizing these specific adaptation rules produce general adaptation rules, that have to be validated by experts.
- Supervised AKA is based on the analysis of adaptations performed by experts. Interviews of experts confronted to decision problems requiring adaptation have been recorded to be afterwards analyzed and modeled within adaptation rules.

Orpailleur is involved in this TCAN project, together with the "laboratoire d'ergonomie du CNAM in Paris", and the Centre Alexis Vautrin in Nancy. Beyond the application framework, this research work will involve progress in the AKA methodology and techniques, that is an original research area in CBR (at its beginning, despite its importance for knowledge-intensive approaches in CBR).

5.3.5. Projects and Collaborations in Spatio-Temporal Reasoning

- Géomatique (CNRS–STIC): “Modélisation, comparaison et interprétation d’organisations territoriales agricoles” (in charge of Florence Le Ber).
- Impact des OGM (MENRT): “Modélisation de la dispersion de transgènes à l’échelle de paysages agricole” (in charge of Florence Le Ber).
- Eau, environnement, sociétés Ressources – Usages – Risques Gestion (CNRS–SHS): RIBAVAL project "Conception d’un outil pour la simulation du fonctionnement d’un bassin versant et définition des conditions d’utilisation pour la co-gestion" (in charge of Florence Le Ber).
- 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.

5.4. Contrat de Plan État-Région “Intelligence Logicielle” (CPER-IL)

- The CPER–IL project ILD-ISTC for “Ingénierie des langues et du document, information scientifique, technique et culturelle”.
  The Orpailleur team is involved within the regional research project ILD-ISTC. In this context, research work is carried out in association with the URI team at INIST CNRS on the design of an operational text mining platform (mainly for technological watch with respect to scientific texts).
- The CPER–IL project BIOINFO for “Bioinformatique et applications à la génomique”.
  The Orpailleur team is involved in three main collaborations with other biology laboratories: namely “Extraction de connaissances pour la compréhension du transfert horizontal chez les bactéries et la dynamique des génomes” (with the laboratory LGM UHP), “Exploitation des génomes – Gènes candidats” (with EA 3446, 3441, 3443 UHP), and “Interactions gène-environnement et maladies cardio-vasculaires” with INSERM U525 (Équipe 4).

6. Dissemination

6.1. 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.

6.2. 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.

7. Bibliography

Books and Monographs

[1] F. LE BER, G. LIGOZAT. *Représentation et raisonnement sur le temps et l'espace – Actes de l'atelier AFIA, Nice*, May 2005.

Doctoral dissertations and Habilitation theses

[2] J.-L. METZGER. *Contribution à l'élaboration d'un modèle de raisonnement à partir de cas pour l'aide à l'interprétation d'organisations spatiales agricoles*, Thèse d'université, Université Henri Poincaré - Nancy 1, Apr 2005.

Articles in refereed journals and book chapters

[3] R. BENDAOUD, Y. TOUSSAINT, A. NAPOLI. *Hiérarchisation des règles d'association en fouille de textes*, in "Revue des Sciences et Technologies de l'Information (Série Ingénierie des Systèmes d'Information)", vol. 1, Jan 2005, p. 263-274.

[4] M. BENOÎT, M. CAPITAINE, F. LE BER. *Méthodes de représentation des règles d'organisation du territoire agricole*, in "Agricultures et territoires", C. LAURENT, P. THINON (editors). , Traité IGAT, série Aménagement et gestion du territoire, vol. 9, Hermès Lavoisier, Apr 2005, p. 191–206.

[5] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *Deux méthodologies de classification de règles d'association pour la fouille de textes*, in "Revue des Nouvelles Technologies de l'Information", Oct 2005.

[6] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *Towards a Text Mining Methodology Using Association Rules Extraction*, in "Soft Computing Journal", Sep 2005.

[7] M. D'AQUIN, C. BOUTHIER, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Knowledge Editing and Maintenance Tools for a Semantic Portal in Oncology*, in "International Journal on Human-Computer Studies", vol. 62, n° 5, May 2005, p. 619-638.

[8] N. JAY, J. FRESSON, M. D. GASPERI, B. LACOUR, D. MAYEUX, F. KOHLER. *Analyse de la trajectoire de soins en cancérologie. Extraction de connaissances à partir du PMSI*, in "Journal d'Economie Médicale", vol. 23, n° 3–4, 2005, p. 191–194.

[9] S. LARDON, F. LE BER, J.-L. METZGER, P.-L. OSTY. *Une démarche et un outil pour modéliser et comparer l'organisation spatiale d'exploitations agricoles*, in "Revue internationale de Géomatique", vol. 15, n° 3, Apr 2005, p. 263-280.

[10] D. LEENHARDT, F. CERNESSON, J.-F. MARI, D. MESMIN. *Anticiper l'assolement pour mieux gérer les ressources en eau : comment valoriser les données d'occupation du sol ?*, in "Ingénieries eau agriculture territoires", n° 42, Jun 2005, p. 13-22.

[11] J.-F. MARI, F. LE BER. *Temporal and Spatial Data Mining with Second-Order Hidden Markov Models*, in "Soft computing", 2005.

Publications in Conferences and Workshops

[12] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *Deux méthodes de classification de règles d'association pour la fouille de textes*, in "12èmes journées de la Société Francophone de Classification - SFC-05, Montréal, Canada", V. MAKARENKOV, G. CUCUMEL, F.-J. LAPOINTE (editors). , Presses Universitaires de Montréal, Apr 2005, p. 104-107.

[13] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Vers une utilisation du critère pessimiste de Wald pour aider à la décision à partir de cas*, in "13ème Atelier Raisonnement à partir de cas, Nice, France", May 2005.

[14] M.-D. DEVIGNES, H. DE PALMA, L. PIERRON, L. DOMENJOUD, M. SMAÏL-TABBONE. *User-designed web services to support heterogeneous biological data retrieval*, in "First International Workshop on workflows management : new abilities for the biological information overflow - NETTAB 2005, Naples, Italy", Oct 2005, p. 27-34.

[15] C. ENG, A. THIBESSARD, S. HERGALANT, J.-F. MARI, P. LEBLOND. *Data Mining Using Hidden Markov Models (HMM2) to Detect Heterogeneities into Bacteria Genomes*, in "Journées Ouvertes Biologie, Informatique et Mathématiques - JOBIM 2005, Lyon, France", Poster, Jul 2005.

[16] S. HERGALANT, B. AIGLE, P. LEBLOND, J.-F. MARI. *Fouille de données du génome à l'aide de modèles de Markov cachés*, in "Extraction et Gestion de Connaissances - EGC 2005, Paris, France", Atelier fouille de données complexes dans un processus d'extraction de connaissances, Jan 2005, p. 141 – 148.

[17] N. JAY, J. FRESSON, M. D. GASPERI, B. LACOUR, D. MAYEUX, F. KOHLER. *Analyse de la trajectoire de soins en cancérologie. Extraction de connaissances à partir du PMSI*, in "Actes des Journées EMOIS 2005", 2005.

[18] R. KASSAB, J.-C. LAMIREL, E. NAUER. *Novelty Detection for Modeling User's Profile*, in "Proceedings of the Eighteenth International Florida Artificial Intelligence Research Society Conference, California", AAAI Press, 2005, p. 830-831.

[19] R. KASSAB, J.-C. LAMIREL, E. NAUER. *Une nouvelle approche pour la modélisation du profil de l'utilisateur dans les systèmes de filtrage d'information basés sur le contenu : le modèle de filtre détecteur de nouveauté*, in "Actes de la COnférence en Recherche d'Information et Applications (CORIA)", 2005.

[20] F. LE BER, M. BENOÎT, M. CAPITAINE, S. LARDON, P.-L. OSTY, C. BRASSAC, J.-M. PRÉAU, L. MONDALA, J.-L. METZGER, A. NAPOLI. *Modélisation, comparaison et interprétation d'organisations spatiales agricoles*, in "Actes du colloque "Société de l'information", Lyon, France", CNRS, Oct 2005, p. 194-198.

[21] F. LE BER, S. LARDON, C. BRASSAC, M. MAINGUENAUD, J.-M. PRÉAU. *Construction collaborative d'objets géo-graphiques*, in "International Conference on Spatial Analysis and GEOmatics - SAGEO 2005, Avignon, France", D. JOSSELIN, T. LIBOUREL (editors). , Actes sur CD, Jun 2005, p. 1-13.

[22] F. LE BER, A. NAPOLI. *Relations, structures et objets : quelques variations*, in "Langages et modèles à objets - LMO'05, Bern, Suisse", M. HUCHARD, S. DUCASSE, O. NIERSTRASZ (editors). , L'objet, vol. 11, n° 1–2, Hermès (Paris), S. Ducasse et O. Nierstrasz, Mar 2005, p. 177–190.

[23] S. MAUMUS, A. NAPOLI, L. SZATHMARY, S. VISVIKIS-SIEST. *Exploitation des données de la cohorte STANISLAS par des techniques de fouille de données numériques et symboliques utilisées seules ou en combinaison*, in "Atelier Fouille de Données Complexes dans un Processus d'Extraction des Connaissances - EGC 2005, Paris, France", Feb 2005, p. 73-76.

[24] S. MAUMUS, A. NAPOLI, L. SZATHMARY, S. VISVIKIS-SIEST. *Fouille de données biomédicales complexes : extraction de règles et de profils génétiques dans le cadre de l'étude du syndrome métabolique*, in "Journées Ouvertes Biologie Informatique Mathématiques - JOBIM 2005, Lyon, France", G. PERRIÈRE, A. GUÉNOCHE, C. GOURGEON (editors). , Jul 2005, p. 169-173.

[25] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Méthode sémantique pour la classification et l'interrogation de sources de données biologiques*, in "EGC 2005 - Atelier Modélisation des Connaissances, Paris, France", Jan 2005.

[26] N. MESSAI, M.-D. DEVIGNES, A. NAPOLI, M. SMAÏL-TABBONE. *Querying a Bioinformatic Data Sources Registry with Concept Lattices*, in "13th International Conference on Conceptual Structures - ICCS 2005, Kassel, Germany", F. DAU, M.-L. MUGNIER, G. STUMME (editors). , Lecture Notes in Computer Science, vol. 3596, Springer, Jul 2005, p. 323-336.

[27] N. MESSAI, M.-D. DEVIGNES, M. SMAÏL-TABBONE, A. NAPOLI. *Treillis de concepts et ontologies pour l'interrogation d'un annuaire de sources de données biologiques (BioRegistry)*, in "Actes du XXIIIème congrès Informatique des Organisations et Systèmes d'information et de décision - INFORSID 2005, Grenoble, France", May 2005.

[28] M. SMAÏL-TABBONE, S. OSMAN, N. MESSAI, A. NAPOLI, M.-D. DEVIGNES. *BioRegistry : a structured metadata repository for bioinformatic databases*, in "First International Symposium on Computational Life Science - CompLife 2005, Konstanz, Germany", M. R. BERTHOLD, R. GLEN, K. DIEDERICHS, O. KOHLBACHER, I. FISCHER (editors). , Lecture Notes in Bioinformatics, vol. 3695, Springer, Sep 2005, p. 46-56.

[29] L. SZATHMARY, A. NAPOLI. *CORON : A Framework for Levelwise Itemset Mining Algorithms*, in "Third International Conference on Formal Concept Analysis - ICFCA '05, Lens, France", B. GANTER, R. GODIN, E. MEPHU NGUIFO (editors). , Supplementary volume Proceedings of The Third International Conference on Formal Concept Analysis (ICFCA '05), Feb 2005, p. 110–113.

[30] S. TÉNIER, A. NAPOLI, X. POLANCO, Y. TOUSSAINT. *Semantic Annotation of webpages*, in "Workshop on Knowledge Markup and Semantic Annotation – SemAnnot 2005, ISWC 2005 Workshop, Galway, Irlande", S. HANDSCHUH (editor). , 2005.

[31] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Decentralized Case-Based Reasoning for the semantic Web*, in "4th International Semantic Web Conference - ISWC 2005, Galway, Irlande", Y. GIL, ET AL. (editors). , Lecture Notes in Computer Science, vol. 3729, Springer, May 2005, p. 142-155.

Internal Reports

[32] H. CHERFI, A. NAPOLI, Y. TOUSSAINT. *Deux méthodes de classification de règles d'association pour la fouille de textes*, Research Report, March 2005.

[33] J.-F. MARI, F. TOUZAIN, S. HERGALANT, I. DEBLED-RENNESSON. *FouDanGA : Fouille de données pour l'annotation de génomes d'actinomycètes*, Poster de présentation de l'ACI FouDanga aux journées JOBIM 2005 (Lyon) : rapport d'avancement à 1 an., Research Report, June 2005.

[34] A. NAPOLI. *A smooth introduction to symbolic methods for knowledge discovery*, Research Report, March 2005.

[35] E. NAUER, A. RICHARD, S. DERRIÈRE, F. GENOVA, A. NAPOLI, Y. TOUSSAINT. *Construction d'une ontologie des descripteurs UCD en astronomie*, Research Report, LORIA, 2005.

[36] A. RICHARD. *Construction d'une ontologie de descripteurs en astronomie à partir de tables de données*, DEA Report, LORIA, 2005.

[37] L. SZATHMARY, A. NAPOLI, S. O. KUZNETSOV. *ZART : A Multifunctional Itemset Miner Algorithm*, Research Report, February 2005.

[38] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Towards a Semantic Portal for Oncology using a Description Logic with Fuzzy Concrete Domains*, Research Report, March 2005.