INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Team Orpailleur

Extraction de connaissances

Lorraine

THEME COG

Activity Report

2004

Table of contents

1. Team

Team Leader

Amedeo Napoli [Research Director CNRS]

Administrative assistant

Christelle Collet [INRIA, from 12/07/2004 to 11/01/2005]
Antoinette Courrier [CNRS, until 12/07/2004]

Staff members

Marie-Dominique Devignes [Research Scientist CNRS]
Florence Le Ber [Professor, ENGEES Strasbourg]
Jean Lieber [Faculty Member, University Henri Poincaré — Nancy 1]
Jean-François Mari [Professor, Nancy II University]
Emmanuel Nauer [Faculty Member, Metz University]
Malika Smaïl [Faculty Member, University Henri Poincaré— Nancy 1]
Yannick Toussaint [Research Scientist INRIA]

Ph.D. Students

Mathieu d'Aquin [Ph.D. Student, MENRT fellowship]
Rokia Bendaoud [Ph.D. Student, INRIA-Région fellowship since October 2004]
Martine Cadot [Ph.D. Student, PRAG, University Henri Poincaré — Nancy 1]
Hacène Cherfi [Ph.D. Student, ATER University Henri Poincaré — Nancy 1, Thesis 15/11/2004]
Adrien Coulet [Ph.D. Student, CIFRE Fellowship with Kika médical since November 2004]
Sébastien Hergalant [Ph.D. Student (co-supervised), INRA-Région fellowship]
Nicolas Jay [Ph.D. Student (co-supervised), Assistant Hospitalier Universitaire à la Faculté de Médecine]
Sandy Maumus [Ph.D. Student (co-supervised), INSERM-Région fellowship]
Nizar Messaï [Ph.D. Student, UHP-Région fellowship since November 2004]
Jean-Luc Metzger [Ph.D. Student, ATER Nancy 2 University]
Frédéric Pennerath [Ph.D. Student since Octobre 2004, Faculty Member of Supélec]
Laszlo Szathmary [Ph.D. Student, KVM-Région fellowship]
Sylvain Tenier [Ph.D. Student, CIFRE felllowship with INIST since April 2004]

Post-doctoral fellows

Rim Al Hulou [Technical staff (contractuelle)]
Huaizhong Kou [Post Doctoral fellowship, ACI MDA from 01-12-2003 to 30/11/2004)]

Visiting scientist

Sergei Kuznetsov [Professor, VINITI Moscow, Russie, du 17/11/2004 au 17/12/2004)]

Technical staff

Sébastien Brachais [Junior Technical staff INRIA]

2. 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. 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 transformation leading from data to knowledge: in the mining operations, in the modeling of the domain for designing a domain ontology (or extending the ontology with extracted knowledge units), and in knowledge representation and reasoning. Finally, the knowledge extraction process and the associated knowledge base can be used for problem-solving and for achieving different tasks 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: Rokia Bendaoud, Martine Cadot, Hacène Cherfi, Sébastien Hergalant, Florence Le Ber, Jean-François Mari, Sandy Maumus, Amedeo Napoli, Frédéric Pennerath, Laszlo Szathmary, Sylvain Tenier, Yannick Toussaint.

knowledge discovery is a process for extracting information units from large databases, units that can be interpreted to become knowledge units to be reused.

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 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 considered data. The analyst selects and interprets a subset of the units for building “models” that will be further considered as knowledge units with a certain plausibility.

The KDD process is performed with a KDD system based on four main components: the databases, a domain ontology (associated with a knowledge-based system), data mining modules (either symbolic or numerical), and interfaces for interactions with the systems, e.g. editing and visualization. A KDD system is aimed at handling huge volume of data in a given domain. For achieving this task, the 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, 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.

Lattice-based classification can be considered as a symbolical data mining technique that can be 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), frequent itemsets i.e. sets of properties (characteristics of data), occurring together with a certain frequency, or association rules with a given confidence (association rules emphasize links between sets of properties). 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 relation between individuals and properties can be read as follows: the individual `i` includes or does not include 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 a parallel way, 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` ⟶ `B` relating the subset of properties `A` with the subset of properties `B`, and 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). This is why a number of measures has been set on, mainly based on probability theory e.g. the so-called *statistical implication* between `A` and `B`, which can be read as `when A almost B`. Moreover, a probability distribution of rules can also be studied. Currently, this kind of work is under investigation, and some results can be found in [13][14][3].

3.1.1.2. Knowledge discovery in chemical reaction databases.

In this section, we briefly present an experiment on knowledge discovery in chemical reaction databases [35][12][11]. 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. The knowledge discovery process relies on frequent levelwise itemset search and association rule extraction, but also on chemical knowledge involved within every step of the knowledge discovery process. Moreover, the overall process is supervised by an expert of the domain. This experiment on mining chemical reaction databases is very original and provides promising results. Moreover, the method for preparing dynamic data to be mined such as chemical reactions can be reused in other contexts where data are under a similar form.

3.1.1.3. 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 which follows every five years apparently healthy French families. At the beginning of the study, in 1993, 1006 families (composed by two parents and at least two chidren) were recruited for medical examination at the Centre for Preventive Medicine (Vandoeuvre-lès-Nancy, France). 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. In other words, there is a need for extracting implicit or new potential risk factors for CVD within an always growing volume of data (mainly due to the development of technologies such as PCR multiplex or microarrays). In the STANISLAS cohort, information hold on environmental, clinical, biological and genetic data. The first experiments gave (1) results in accordance with the domain knowledge and (2) new results giving new research insghts for further investigations [21].

With respect to statistical work more generally used in such a context, the general idea of this research work is to mine the cohort for extracting itemsets that are in turn considered as hypotheses to validate by statistical tests.

More specifically, this year, we performed experiments concerning a specific pathology related to CVD, namely the metabolic syndrome (MS), which is a cluster of CV risk factors that some people of the

STANISLAS cohort have. We have made experiments with an open source software named WEKA. Briefly, we used mostly the J48 algorithm which is the implementation of the famous C4.5 decision tree algorithm. The first results are very encouraging and gave precious information on MS in the cohort. In parallel, for studying MS in the cohort, we are currently doing first tests with J-Close (a Java implementation of the Close algorithm, see § 4.1) that extracts closed frequent itemsets and association rules.

In the next future, we project to combine the two methods described above that are symbolic and numeric data mining methods, in order to design a new technique allowing to analyze complex data such as those of the STANISLAS cohort.

3.1.2. Hidden Markov Models for Data Mining

We present in this section the research work based on higher-order stochastic models – namely second-order hidden Markov models (HMM2) – that aims at discovering spatial and temporal dependencies in databases. HMM2 are able to map sequences of data into a Markov chain in which the transitions between the states depend on the *two* preceding states. HMM2 are based on probability and statistics theories. Their main advantage 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 points: (1) The elaboration of a process for mining spatial and temporal dependencies for knowledge acquisition. This process involves a non-supervised classification of data. (2) The specification of adequate visualization tools giving a synthetic view of the classification results to the experts, who have to interpret the classes and/or specify new experiments.

Below, we describe three main applications, developed using a generic data-mining system for spatio-temporal data, based on HMM2, and named CARROTAGE (the CARROTAGE system is a free software with a GPL license). The two first experiments concern with knowledge discovery in the domain of agronomy (done in collaboration with agronomists), one for a better understanding of the farmer work, and the other for a better management and prediction of water needs The third experiment is about gene segmentation and interpretation in the domain of bioinformatics.

3.1.2.1. Crop rotations in the Seine river watershed.

For thirty or forty years, the hydrosystem of the Seine river has been gradually degraded, regarding water quality and biological population, due to human activities (domestic, industrial, agricultural activities). The nitrate contamination of cave and surface waters is mainly caused by the evolution of agricultural activities, and related to their nature and to their organization inside the river watershed. The objective of the interdisciplinary research program PIREN-Seine (*Programme Interdisciplinaire de Recherche en ENvironnement sur la Seine*) is to develop a tool for predicting the water quality in the Seine river watershed, based on assumptions on agricultural changes. In this research work, members of the INRA (“Institut de la recherche en agronomie”) team in Mirecourt analyze the agricultural activities in the watershed, with respect to their dynamics and their spatial organizations. They particularly focus on the crop (temporal) rotations that may explain the risk of nitrate dissemination. In this application we use a French national database related to land use, named `Ter Uti`, that describes the land use at two levels: the first one is defined by a grid of aerial pictures, and the second level is defined by 6x6 matrices of sites located in the pictures. Land use (wheat, corn, forest,...) is checked every year on each site.

HMM2 have been used for computing the average crop distribution during a given time period (here from 1992 to 1999), for viewing the main annual transitions between crops, and for listing all types of crop rotations in each region. Such an analysis has been carried out for small agricultural regions in the Seine watershed. The regions are then clustered according to their main crop rotations and their evolutions. This classification has appeared to yield meaningful results for domain experts, especially for specifying simulation models of nitrate dissemination.

3.1.2.2. Interpretation of satellite images of the Midi-Pyrénées Region.

Our approach has been used by researchers of the INRA research center in Toulouse that works on the prediction of irrigation needs in the Midi-Pyrénées region (South-West of France). Usually, irrigation needs are estimated using annual land-use maps based on satellite data. This method is not always satisfactory since data are not necessarily available at the moment the prediction has to be done: the satellite images are obtained at the beginning of the cropping season (in spring) and this does not allow to recognize all the crops of a given region. Whenever the crop rotations are known, they can be used for recognizing the crops themselves, based on the land-use map of the year before. Knowing the crop in a plot at year $n - 1$, the potential crops in the same plot at year n can be inferred, and their number reduced using the available satellite images.

We perform a spatial clustering by defining a fractal scanning of the images with the help of a Hilbert-Peano curve, that introduces a total order on the sites, preserving the relation of neighborhood between the sites. Spatial and temporal classifications are simultaneously processed by means of two HMM2 measuring the *a posteriori* probabilities to map a temporal sequence of images onto a set of hidden states. In this case, we have adopted a Bayesian point of view for measuring the uncertainty of a classification by a probability.

The models built for the spatial segmentation and the temporal segmentation have been used in two complementary ways. The first one allows the definition of homogeneous and stable areas regarding the crop rotations, while the second one allows a more specific study of each area. This method has appeared to be very interesting, although the scale of `Ter Uti` data may be insufficient for precisely recognizing the potential crops in an irrigation basin.

3.1.2.3. An application in bioinformatics.

A long-term data mining research project in bioinformatics is carried out in collaboration with the Laboratory of Genetics and Microbiology of the “Université Henri Poincaré Nancy 1” (thesis of Sébastien Hergalant) [17]. In April 2004, this project has been selected in the ACI ImpBio.

The biological material is the soil-dwelling filamentous bacteria belonging to the genus *Streptomyces*, that is the largest source of antibiotics amongst microorganisms. In particular, the *Streptomyces coelicolor* chromosome (8,7M bases) is entirely sequenced and annotated. We are interested in detecting “genome heterogeneity islands”, and inter-sequences dependencies by means of Hidden Markov Models, without prior knowledge. Initially, we have focused on the understanding of horizontal transfer phenomena, but two other areas of interest in genetics are currently investigated: the detection of intragenomic DNA repetitions, and the detection of promoters. So far, the mining of the DNA sequences is performed under the supervision of Sébastien Hergalant, who is specifying a toolbox for computing, displaying and analyzing the *a posteriori* probabilities of the HMM2 hidden states, in various adapted graphical standard environments. It then becomes possible to correlate the output signal of HMM2 with the biological annotations of long segmented DNA sequences (more than 50 000 nucleotides).

We are investigating the pertinence of various methods for segmenting and classifying the chromosome. At present, we are using a *Fast Fourier Transform* to extract the periodic components, and to have another point of view on the stationary/transient behaviors of the process. We also try to determine an appropriate distance, e.g. the Mahalanobis distance, between a codon in a gene, and the class of codons specific to the species.

Moreover, we have elaborated several learning methods for specific analyses:

- Understanding of horizontal transfer understanding. Markovian models with respect to “species specific homogeneities” have been designed and coupled with the transform filters described above. Their behavior generates regions with different statistical properties allowing the user to separate “foreign DNA regions” with the proper DNA regions of the studied species. In *S. coelicolor*, the regions with statistical consensus have been extracted and correlated with potential events of horizontal transfer.
- Detection of intragenomic reiteration detection. In previous work, we have developed Markov models for processing DNA sequences without prior knowledge of their content, searching for DNA subsequences that show strong homology. The graphical signal was used to detect different forms of repeats, e.g. tandem, direct or reversed, with various length and localization in the *S. coelicolor* genome. The variability captured during the learning step by the EM algorithm allows the detection of degenerated repetitions. This interesting feature is not present in general in the string algorithms searching for exact matches.

- Detection of promoters. We have studied the detection of short DNA patterns in the chromosome, appearing frequently (and abnormally) at non-random locations. Identifying the underlying functions linked to these short DNA patterns may be of great interest in the decryption of genome organization and regulatory functions. First DNA patterns identified with the present method have been identified as promoters (promoters are regulatory sequences essential for the cell life, constituting target sites for specific proteins and implied in the gene expression mechanisms). We are currently working on the automatic extraction of these patterns. The resulting set of classes, defined by the underlying DNA sequences, could characterize new promoters and thus define new sets of co-regulated genes.

3.1.3. Text Mining

The goal of a text mining process is to find new and useful knowledge units in a large set of texts. If text mining relies on the principles of KDD, it shows 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 discovery process more complex. To avoid information dispersion, a text mining process has to take into account paraphrases, ambiguities, specialized vocabulary, and terminology. Moreover, 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 should not be extracted by the mining process as new knowledge. Part of it is not expressed but may be 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: actual texts in actual contexts with robust tools, contrasting other works dealing with specific phenomena in the language. The language is considered as the way to access information, and not an object to be studied *per se*. Thus, the text mining process is considered as involved in a loop, where the process can be used to improve linguistic resources, and where linguistic resources can be used, in turn, to improve the information extraction process for guiding a kind of model-based text mining process (the model makes reference to the available knowledge on the domain of texts).

3.1.3.1. The process of text mining.

The expression “text mining” is widely used in the literature to name very different experiments, starting from information retrieval or question answering to ontology building or technological watch [8]. >From our point of view, we define text mining as a specific process of knowledge discovery in databases. An analyst, expert in a scientific or technical domain, is in charge of guiding the mining process of a large amount of texts. The very first steps are dedicated to linguistic knowledge acquisition: lexicon, terminology, markers of semantic relations, discourse markers, specific syntactic or semantic structures...The following steps aim at identifying or structuring the background knowledge for extracting new knowledge units.

3.1.3.2. Extraction of association rules from texts.

We have performed a number of experiments on the extraction of association rules from texts, 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.

- Ranking association rules. Texts are indexed using a merge of several thesauri (we are currently working on medical texts). The support and the confidence are the two main indices 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 indices, together with other indices, e.g. the interest, the dependency, the novelty have been studied. The rules, ranked according to these indices, have been presented to the analyst. It turns out that some combinations of these indices to rank the association rules allow the analyst to identify complex semantic relations between terms or synonyms. We thus propose a new algorithm to combine these indices. However, most of the extracted rules are in accordance with the present domain knowledge, and do not provide any new knowledge unit. Thus, the next objective is to be able to extract association rules providing new effective knowledge units.

- Introducing a knowledge model. The previous experience on statistical indices leads us to adopt a strong hypothesis: text mining should be performed in accordance to an existing knowledge model. Text mining is then considered as an interactive and incremental process where the knowledge base is used to rank the association rules and where the association rules are used to increment the knowledge base. We defined a likelihood measure which rank the association rules, with respect to a "degree of novelty" that a rule may include, compared to the existing background knowledge [18][15][1]

3.1.3.3. Structuring association rules into hierarchies

This work aims at structuring associations into hierarchies at two different levels. At the global level, subsumption between rules follows subsumption into the Galois lattice of the close itemsets which are used to extract the association rules. At a local level, we take into account a knowledge model which structures properties into a hierarchy. This knowledge model is used to build for each rule a generalisation hierarchy of association rules [34].

3.2. Knowledge Representation and Knowledge Systems

Keywords: *case-based reasoning, classification-based reasoning, description logics, knowledge representation, lattice-based classification, object-based representation systems, qualitative spatial reasoning*.

Participants: Mathieu d'Aquin, Sébastien Brachais, Florence Le Ber, Jean Lieber, Jean-Luc Metzger, Amedeo Napoli, Laszlo Szathmary.

knowledge representation is a process for representing knowledge within knowledge representation formalisms, giving knowledge units a syntax and a semantics.

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 an associated 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 also 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 object-based knowledge representation (OBKR) systems, and description logic (DL) systems, together with classification-based reasoning. The function of such a system is to represent knowledge units within concepts (also called classes), attributes (that can be 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. Such a system provides a representation and an organization of knowledge units, and a number of inference services. Among the inference services, let us mention concept and individual classification. The first operation is used to insert a concept at the right place in the concept hierarchy (searching for its most specific subsumers and its most general subsumees). The second operation 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" (a recent overview of problems and systems is proposed in [7]).

Case-based reasoning (CBR) 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 concept hierarchy, case-based reasoning can be seen as a natural extension of classification-based reasoning. Retrieval and adaptation may be based both on classification and on searching for paths in the concept hierarchy. 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 in the next section). These kinds of 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. 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, CBR. This research work is applied to answer agronomical questions regarding the recognition and the analysis of farmland spatial structures.

3.2.2.1. Lattice-based classification of spatial relations.

This work was initiated during the thesis of Ludmila Mangelinck (1995–98) 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 The main objective of this system, called LOLA, is to recognize *landscape models* on land-use maps extracted from satellite images. Landscape models are abstract models describing agricultural spatial structures as sets of spatial entities and qualitative spatial relations between these entities. They are used to classify *zones* extracted from the maps. A zone is a collection of raster regions, i.e. connected sets of pixels with the same label denoting the land-use category, e.g. crops, meadows, forest, buildings, etc. >From an implementation point of view, an object-based knowledge representation system, equipped with a classification process, has been used. In this framework, the exploitation of land-use maps for landscape analysis may be considered as an *instance classification* problem, where landscape models correspond to classes, while zones correspond to instances that have to be classified according to landscape model classes.

Following these needs, 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.

3.2.2.2. CBR on spatial organization graphs.

This work has been undertaken in the framework of J.-L. Metzger's thesis in collaboration with INRA SAD and ENGREF. The objective is to develop a knowledge-based system, called ROSA, for comparing and analyzing farmland spatial structures. The reasoning in the ROSA system follows the principles of case-based reasoning (CBR). In our research work, CBR relies on the agronomical 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.

- 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 vertices 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) [6].

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. Finally J.-L. Metzger will defend his thesis in the beginning of 2005.

3.2.3. Knowledge Management in Medicine: the Kasimir System

This section presents an overview of the KASIMIR research project, whose objective 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 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 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 protocol case, oncologists try to *adapt* the protocol (actually they discuss such a case during meetings of the so-called "breast therapeutic decision committee", including experts of all domains in breast oncology, e.g. chemotherapy, radiotherapy and surgery). In addition, protocol adaptations are currently 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 of 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 some knowledge. The adaptation knowledge acquisition (AKA [20]) is a current research work which takes three directions: the AKA from experts, the automatic AKA and the semi-automatic AKA.

AKA from experts consists in analysing adaptations performed by experts. Interviews of experts confronted to decision problems requiring adaptation have been recorded to be afterwards analyzed and modeled within adaptation schemas.

Automatic AKA is based on the mining of the protocols. A protocol can be seen as a set of rules `situation`⟶`decision`. Knowing how the decisions change when the situations change from one rule to another rule provides a specific adaptation rule. By generalisation of these specific rules, general adaptation rules are obtained. This generalisation has been implemented thanks to a frequent close itemset extraction tool. This requires a formatting of the situations and decisions of the protocol by itemsets [30]. A system, called CABAMAKA, realises 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 schemas that are intelligible but cannot be directly operationalised, while the latter provides adaptation rules that can be directly implemented but are difficult to understand. The aim of the semi-automatic AKA will be to combine these two kinds of AKA in order to obtain operational *and* intelligible adaptation knowledge.

The research in AKA is carried out in the interdisciplinary project TCAN (see section 5.2.4).

3.2.3.2. Knowledge representation for decision support tools.

The KASIMIR system is currently under development and implements (at the moment) an object-based representation formalism associated with an inference engine based on hierarchical classification, and a decision support module in oncology. 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 propose more than one treatment for “borderline cases” [27][28]. A study of formalisms such as fuzzy description logics in which such inferences can be made has also been carried out [32]. Another study is about multiple viewpoint representation and reasoning, which is useful for the modelling of reasoning of the breast therapeutic decision committee (each viewpoint represents a domain in breast oncology). In [31], a multiple viewpoint model has been presented. It is currently extended in order to take into account more interactions between viewpoints.

3.2.3.3. Going further: 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, which is a formalism adapted to the semantic Web. A script for translating decision protocols from the *ad hoc* formalism to OWL has been written [29]. Currently, the representation of protocols is also re-engineered in order to better benefit from OWL expressiveness.

This work also implies a new software architecture, for the KASIMIR reasoner and the editing, visualization and maintenance modules. This architecture must take into account constraints related to the distributed and dynamic environment of the semantic Web. In order to interrogate the protocols written in OWL, an instance editor called EDHIBOU has been developed. 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 developped for this purpose.

3.3. The Semantic Web

Keywords: *Semantic Web, bioinformatics, information retrieval guided by data mining, knowledge-based information access and retrieval.*

Participants: Rokia Bendaoud, Adrien Coulet, Rim Al Hulou, Mathieu d'Aquin, Marie-Dominique Devignes, Huaizhong Kou, Nizar Messaï, Amedeo Napoli, Emmanuel Nauer, Malika Smaïl, Laszlo Szathmary, Yannick Toussaint.

Semantic Web is a framework for building knowledge-based systems for manipulating documents on the Web by their contents and their semantics.

3.3.1. 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, 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 will be able to read, understand, and manipulate information on the Web only if the knowledge necessary for achieving those tasks is available. This is why ontologies are of main importance with respect to the task of building a semantic Web. Moreover, there is a need for languages for annotating documents, i.e. describing the content of documents, and giving a semantics to this content. Knowledge representation languages are good candidates for achieving this kind of task: they have a syntax with an associated semantics, and thus they can be used in information retrieval, query answering, and reasoning processes [7].

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. Among others, we are interested in the content-based manipulation of textual documents using annotation, ontologies, and a knowledge representation language. The idea here is to build an XML-based "bridge" between documents and object-based knowledge units associated to the domain of documents. The annotations attached to documents and the queries are built with the help of a domain ontology, and have an XML syntax. Moreover, annotations, elements of the ontology, and queries, have corresponding elements within the knowledge representation language, in our case description logics, providing a semantics for these elements. Then, the manipulation of annotations, e.g. information retrieval, query answering, is left to the reasoning module associated to the knowledge representation formalism.

3.3.2. Intelligent Access to Information

The availability and retrieval of information is of main importance in scientific and technical domains, for e.g. research and watch purposes. Nowadays, there is a huge quantity of data available, and this requires to implement adapted tools for exploiting and taking advantage of the data. One research work carried out within the Orpailleur team concerns the definition and implementation of an environment allowing an intelligent access to information, by combining information retrieval, hypertext navigation, and data-mining. This environment can be used for document retrieval on the Web, bibliographical search or domain analysis.

In this way, we are currently working on the design of a semantic-based approach for comparing and classifying documents. In this approach, the annotations of documents are considered as labelled trees where nodes and edges represent concepts 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 [9][10][33].

Another important and underlying idea 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. 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.

Moreover, information retrieval may be improved when semantic relationships between textual data are taken into account, by making precise the context knowledge of a query, and by filtering documents on the basis of the similarity of their content, with respect to background knowledge. These are the two main ways on which relies our research work for identifying relevant documents on the Web, and in bibliographical databases.

In a more concrete way, in the domain of bibliographical data, the knowledge units extracted from bibliographical data can be used to guide a keyword-based document retrieval on the Web. We have made a number of experiments in a knowledge management perspective, holding on the bibliographical items of the Orpailleur team [22][23][25][24][26]. We have used lattice-based classification with the help of a domain ontology. The lattice-based classification allows to organize the bibliographical items according to the research themes or any other useful or interesting information that is under study. In this way, it is possible to analyze the global work of the team, e.g. to discover who is working with whom, which persons are working on related topics, or which documents are written about these related topics.

3.3.3. Intelligent Access to Bioinformatics Data Sources

Web data sources are widely used in bioinformatics. Scientists are getting more and more concerned with the problem of exploiting at best the mass of biological information stored in the numereous and heterogeneous biological public databases. More than 500 such databases have been listed recently [38], reflecting the complexity of the various biological objects concerned: genes, proteins, transcripts, metabolic pathways, mutations, diseases, and the diversity of living organisms studied: bacteria, plants, animals, model organisms, etc.

Answering a question in biology very often starts with querying one or more data sources. For example, a simple question such as : “what are the human genes from chromosome X that are preferentially expressed in brain” leads to query a human genome data server such as UCSC Genome Browser, EBI Ensembl or NCBI Mapviewer and one or more expression databases. Various integrated systems exist today that can be grouped in two categories. The first category proposes a unified access to a subset of web sources. Portals (PBIL, Infobiogen, Bioweb) or systems such as SRS, Entrez (involving a specific query language) are unified interfaces to a large number of heterogeneous sources and may provide some assistance in selecting appropriate sources for given queries. The second category of systems involves a unified data model for querying heterogeneous sources. Mediation systems such as TAMBIS or DiscoveryLink allow in certain cases automatic processing of complex queries. Data warehouses such as GUS or GenExpress can be constructed to robustly handle such queries over a small number of sources. The systems belonging to the first category (unified access) have a very broad coverage in terms of data sources but a rather low integration power with respect to distributed queries. On the contrary the systems from the second category (unified data model) are very powerful with respect to integration of answers to complex queries but cover a limited number of sources and sometimes present a very limited diffusion. In the semantic web, the challenge is to increase the integration power for an open set of data sources.

Our approach within the Orpailleur team is based on the distinction between two types of problems generated by complex questions in biology: first, the selection of relevant data sources, second, the integration of data from the selected sources. Grouping both aspects in a single system is our ultimate goal and should be achieved by a mediation architecture. However, biological data and data sources are so complex that it has appeared reasonable to divide the work in two parts.

Previous work addressed the second problem and resulted in Xcollect software (presented in section 4.8) which gives a generic solution for automated collecting and integration of biological data given a user-defined scenario. We have also investigated the problem of “homologous” answers retrieved from several sources (e.g.

different functional annotations for a given gene retrieved in different data sources) and how the variation in quality between sources (e.g. update frequency, manual revision) must be taken into account when presenting the answers to the user [16].

Current work deals with formalisation and exploitation of knowledge about data sources towards more efficiency and accuracy in the discovery of relevant sources. The goal is to build a biological registry called BioRegistry gathering together appropriate metadata about web data sources. A survey of various existing resources (Dublin Core Standard, FGDC, DBCat...) led us to distinguish four metadata categories : identification, content, quality and access and availability [37]. Moreover, the BioRegistry also includes metadata tracking and the description of relationships between sources. Well known biological ontologies are used to assign values to content metadata. At this stage of the work, an XML schema has been designed to implement the BioRegistry model that allows using any available ontology. The first instance documents have been created manually and involve MeSH and NCBI taxonomy as ontologies. In the future, feeding the BioRegistry with information about novel sources could benefit from web mining procedures. First exploitation of the BioRegistry is form-based querying, triggering structured information retrieval of the metadata. This should allow the biologist to formulate a multi-criteria query combining various metadata categories and to recover a sorted list of data sources. However, this method does not offer an overall view of the BioRegistry that would allow for browsing during the process of source discovery. An approach for this problem is being studied using formal concept analysis (FCA): Galois lattices are constructed on the basis of binary source properties extracted from BioRegistry resulting in hierarchical source organisation reflecting property sharing between data sources [36]. The user query is first inserted in the hierarchy, thus allowing for subsequent navigation in the neighbourhood of the best fitting sources. Furthermore, this original FCA application is extended by introducing an ontology-based enrichment of user-query in order to take into account semantic relationship between certain properties [36].

4. Software

4.1. Software for Data Mining

Keywords: *association rule extraction, closed frequent itemset search, data mining, frequent itemset search*.

Participant: Laszlo Szathmary.

One of the goals of data mining is to extract hidden relations among objects in databases. Usually frequent itemsets are used to find out association rules, but a huge number of rules is produced, leading to the associated problem of “mining the set of produced rules”. Some recent studies have shown that it may be interesting to find only a subset of particular frequent itemsets, called *closed frequent itemsets* (FCIs). In turn, FCIs can also be used for finding useful association rules.

We have developed several programs for data mining listed hereafter. J-CLOSE is an effective Java implementation of the *Close* algorithm, and is used for extracting FCIs from arbitrary binary contexts. ASSRULEX (“Association Rule eXtractor”) is another software based upon J-CLOSE, which can be regarded as a framework for rule mining algorithms. Both programs are implemented in Java, and thus can be easily reused in other Java applications. In particular, J-CLOSE is used in CABAMAKA, a part of the KASIMIR system. Finally, the *Titanic* algorithm has been implemented too (J-TITANIC), that can be used for finding FCIs and for building iceberg concept lattices. Actually, J-TITANIC is integrated in the GALICIA 2.0 platform.

4.2. Stochastic Systems for Knowledge Discovery

Keywords: *Hidden Markov models, stochastic process*.

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

4.2.1. Carrotage

One aspect of data-mining is to provide a synthetic representation of data that a domain analyst can interpret. The purpose of the CARROTAGE system is to build a partition—called the hidden partition—in which the inherent noise of the data is withdrawn as much as possible [19]. 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.

CARROTAGE is a free software under a GPL license, that takes as input an array of discrete data (the rows represent the spatial sites and the columns the time slots), and that builds 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, in order to build models to simulate the nitrate contamination of cave and surface waters.

4.2.2. genExp

GenExp is an experimental software that has been developed in the framework of the "OGM Impact project", in collaboration with biologists of ESE UPRESA 8079 CNRS, Paris Sud. The objective of the system is to simulate agricultural landscape for studying the dissemination of vegetal transgenes. The system is based on the stochastic system CARROTAGE, and on computational geometry.

4.3. Software for Text Mining

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

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

We are currently developing a system named RAR, standing for "Ranking Association Rules", that allows for navigation through a large set of association rules (such as those obtained within a text mining experiment). This 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 data analyzed. The association rules are supposed to be extracted by a mining algorithm—the *Close* algorithm in our case, for extracting frequent itemsets and association rules—and should be encoded in a predefined XML format. The RAR system then stores the rules in a database, and proposes eight different statistical measures, e.g. support, confidence, interest, conviction, dependence... for sorting the analyzed set of rules. It is also possible for the analyst to focus on smaller sets of rules satisfying a given set of constraints. These constraints may be expressed as 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, Jean-Luc Metzger [contact person].

Rosa, for "Reasoning on Organization of Space in Agriculture", is a system developed in collaboration with agronomists, whose objective is to record and maintain an agronomical 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) that link spatial entities with spatial relations. A vertex of a SOG (either spatial entity or relation) is labeled and linked to a concept of the domain knowledge hierarchy. The functioning of farms is described within "explanations" attached to parts of SOGs. An explanation concern a particular function in the 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, object-based representation system*.

Participants: Mathieu d'Aquin, Christophe Bouthier [ECOO research project], Sébastien Brachais, 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 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 from the protocol being edited: PALÉTUVIER and HYPERTREE (currently developed in the ECOO team at LORIA). 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.

4.6. Intelligent Access to Information for the Semantic Web

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

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

Two major systems are under development. A first one, called "IntoBib", is a generic system designed for the exploitation of bibliographical data. 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 (with respect to a given point of view), normalization of fields e.g. author name, keywords.

A second system is under development for manipulating textual documents by their content. Documents are annotated using a domain ontology (experiments have been carried out with biological documents). The annotations and the ontology are implemented within an XML-based description language. Every element with an XML-based description, i.e. a concept of the ontology or an annotation, has a corresponding knowledge structure within the associated knowledge representation system, currently based on the the RACER description logic system. In this framework, syntactic aspects are taken into account within the XML level, and the semantic aspects are taken into account within the knowledge representation level. The system is aimed at content-based document retrieval, and query management, e.g. query answering, query classification, detection of similarities on the content of documents.

4.7. DefineCrawler: a Generic Crawler for the Semantic Web

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

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

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 the one hand on a thorough study of the capabilities provided by classical information retrieval architectures, and on the other hand on the 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...

4.8. From Xcollect to web services

Keywords: *bioinformatics*, *data integration*, *generic solution*, *user-defined scenario*, *web service*.

Participants: Herve de Palma, Marie-Dominique Devignes [contact person], Malika Smaïl.

The Xcollect project is aimed at managing query answering scenari in biological data sources. The Xcollect application (described in former reports) offers two main functionalities: first, an assistance in the design of a retrieval scenario involving relevant sources, second, the enactment of this scenario to collect and integrate desired data.

Several scenarios have been implemented such as the Xprom scenario used in the collaboration with Lionel Domenjoud [4] and described in last year report. The Xfunction scenario, which retrieves from various data sources all possible functional data associated to a given transcribed sequence, has been used to explore the management of multiple answers to queries [16]. The PromLoc scenario has been designed to tackle with retrieving human gene promoter regions based on first exon identification in sequence databases. Previous work on full-length cDNA characterization [2] and a collaboration with Nadine Martinet at CRB (Centre de Ressources Biologiques, CHU Nancy) have revealed the diversity and complexity of gene structures including multiple promoter usage for a given gene. The Xcollect scenario model could not handle all possible situations. The PromLoc application is being developed that includes the enactment of several elementary Xcollect scenarios. The Xmap scenario was previously implemented as a dedicated application. It was recently used in a large collaborative work which aimed at annotating a large collection of full-length cDNAs [5]. A new version of the Xmap application is currently under development using Xcollect.

Web services have been deployed using java technology (WSDP, JAX-RPC, etc.) to favor the diffusion of Xcollect and instantiated scenarios.

5. Other Grants and Activities

5.1. The European Network of Excellence Knowledge Web

Here is the abstract of the Knowledge Web proposal that has become in 2004 a European network of excellence (three INRIA teams are involved in Knowledge Web: ACACIA at INRIA-SOPHIA, EXMO at INRIA-RHÔNE-ALPES and Orpailleur). The current World Wide Web (WWW) is, by its function, the syntactic Web where structure of the content has been presented while the content 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 the added semantic information in a machine-understandable and machine-processible fashion. It 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.

Spoken in a nutshell, 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. National initiatives

5.2.1. ACI IMPBIO: the FouDAnGA project

Our FouDAnGA proposal (Fouille de données pour l'annotation de génomes d'actinomycètes) to the bioinformatics ACI IMPBIO has been selected in June 2004. This project involves two research teams from Loria and the Lab. of Genetics and Microbiology of the University UHP-Nancy 1. For almost three years, these three groups have been maintaining a tight interaction in the CPER "Intelligence logicielle – Bioinformatics and applications to genomics". This ACI has reinforced and structured our project. This allows two students in co-supervision to complete their second year of thesis.

The scientific motivation of this ACI is to make emerge DNA under-sequences with informative and significant values in molecular genetics, in particular we are studying the signals implied in the genes's regulation.

The models used correspond to the bacteria of the group of the actinomycetes —in particular to Streptomyces— the principal producer of antibiotics and of metabolites with therapeutic interest, and with Mycobacteries (for example *M. tuberculosis*) which 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 [17].

5.2.2. ACI IMPBIO: the ISIBIO project

The ISIBIO project (Information Systems Integration in Biology) is supported since July 2004 by the Ministry of Research in the framework of the ACI IMPBio initiative. This interdisciplinary working group is interested in exploring the role of metadata and ontologies in information systems integration in biology. This project will reinforce the existing collaborations and stimulate new interactions at both national and international levels by the means of organizing twice a year an international seminar. It is also expected that the international visibility of the French commmunity in this domain will benefit from this animation activity.

5.2.3. ACI "Masse de données in Astronomy"

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. A step further must be done at present, and interested researchers should have access to the content of documents, e.g. journal articles, astronomical object catalogs, or measure tables. 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 within the Orpailleur team concern the study and the design of an ontology for representing astronomical objects, starting from a collection of articles (and thus involving text mining) and extending the UCD database. 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.

5.2.4. 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 "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` ⟶ `decision`. Knowing how the decisions change when the situations change from one rule to another rule provides a specific adaptation rule. Clustering and generalizing these specific 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" and the Centre Alexis Vautrin. Beyond the application framework, this work should involve progress in AKA methodology and techniques, that is an original research area in CBR (at its very first beginning, despite its importance for knowledge-intensive approaches in CBR). A preliminary study has been carried out in, that has highlighted several adaptation schemas that remain to be instantiated.

5.2.5. Projects and Collaborations in Spatio-Temporal Reasoning

- Géomatique (CNRS–STIC): "Modélisation, comparaison et interprétation d'organisations territoriales agricoles" (in charge of F. Le Ber).
- Impact of GMO (MENRT): "Modélisation de la dispersion de transgènes à l'échelle de paysages agricole" (in charge of F. Le Ber).
- Other collaborations: INRA (Nancy-Mirecourt, Paris-Grignon, Dijon, Toulouse), Laboratoire ESE UPRESA 8079 CNRS/Paris-Sud, Équipe Codisant, LPI GRC, Université de Nancy 2, GRIC UMR 5612 CNRS Lyon, ENGREF Clermont-Ferrand.

5.2.6. Other Links with CNRS

- AS “Fouille de textes” (Text Mining), and AS Discovery Challenge.
- RTP 12 : “Information et connaissance : découvrir et résumer”.
- AS 127 : “Intégration et Interopérabilité de sources de données génomiques”, attached to RTP 41 “Bioinformatique : de la séquence génomique à la fonction biologique”.
- Working group Ontologies and Metadata for Biology depending on the IMPG action “Informatique, Mathématiques et Physique pour la Génomique”.

5.3. Le Contrat de Plan État-Région (CPER) Intelligence Logicielle

- Project ILD-ISTC (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 done in association with the URI team at INIST CNRS on the design of an operational text mining platform for technological watch.
- Project “Bioinformatique et applications à la génomique”. The orpailleur team is involved in three operations supported by this project, in collaboration with three biology laboratories : Extraction de connaissances pour la compréhenion du transfert horizontal chez les bactéries et la dynamique des génomes (with LGM, UMR UHP-INRA 1128) ; Exploitation des génomes - Gènes candidats (with EA3446, UHP) ; Interactions gène-environnement et maladies cardio-vasculaires (with INSERM U525 - Equipe 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, mainly in “Actions spécifiques du CNRS” as mentioned above.
- 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

Doctoral dissertations and Habilitation theses

[1] H. Cherfi. *Étude et réalisation d'un système d'extraction de connaissances à partir de textes*, Thèse d'université, UHP - Nancy 1, November 2004, http://www.loria.fr/publications/2004/A04-T-532/A04-T-532.ps.

Articles in referred journals and book chapters

[2] O. Bertaux, E. Toselli-Mollereau, C. Auffray, M.-D. Devignes. *Alternative usage of 5' exons in the chicken nerve growth factor gene: refined characterization of a weakly expressed gene*, in "Gene", vol. 334, Jun 2004, p. 83–97.

[3] M. Cadot, J. di Martino. *A Data Cleaning Solution by Perl Scripts for the KDD Cup 2003 Task 2*, in "SIGKDD Explorations", vol. 5, n^{o} 2, January 2004, p. 154–155.

[4] P. Collet, L. Domenjoud, M.-D. Devignes, H. Murad, H. Schohn, M. Dauça. *The human Semaphorin 6B gene is down regulated by PPARs*, in "Genomics", vol. 83, n^{o} 6, Jun 2004, p. 1141–1150.

[5] T. Imanishi, M.-D. Devignes, S. Sugano, e. al.. *Integrative annotation of 21,037 human genes validated by full-length cDNA clones.*, in "PLoS Biology", on-line publication, vol. 2, n^{o} 6, April 2004, p. 856–875, http://www.loria.fr/publications/2004/A04-R-517/A04-R-517.ps.

[6] F. Le Ber, C. Brassac, J.-M. Préau, J.-L. Metzger. *De la confiserie sur le Causse, ou comment concilier chorèmes et graphes*, in "Agro-tribulations", C. Blanc-Pamard, J.-P. Deffontaines, S. Lardon, C. Raichon, S. Zasser-Bedoya (editors)., INRA éditions, September 2004, p. 117–129.

[7] A. Napoli, B. Carré, R. Ducournau, J. Euzenat, F. Rechenmann. *Objets et représentation, un couple en devenir*, in "L'objet", vol. 10, December 2004, p. 61–81.

[8] Y. Toussaint. *Extraction de connaissances à partir de textes structurés*, in "Document numérique", vol. 8, n^{o} 3, December 2004, p. 11–34.

Publications in Conferences and Workshops

[9] R. Al Hulou, A. Napoli. *Utilisation de connaissances pour l'aide à la recherche documentaire fondée sur le contenu*, in "4èmes Journées d'Extraction et de Gestion des Connaissances - EGC'2004, Clermont Ferrand, France", G. Hébrail, L. Lebart, J.-M. Petit (editors)., Poster, RNTI - Cépaduès Editions Toulouse, January 2004, 503, http://www.loria.fr/publications/2004/A04-R-018/A04-R-018.ps.

[10] R. Al Hulou, A. Napoli, E. Nauer. *Une mesure de similarité sémantique pour raisonner sur des documents*, in "Langages et Modèles à Objets - LMO'04, Lille, France", J. Euzenat, B. Carré (editors)., Numéro spécial L'objet, vol. 10, n^{o} 2-3, Hermès, Paris, March 2004, p. 217–230.

[11] S. Berasaluce, C. Laurenço, A. Napoli, G. Niel. *An Experiment on Knowledge Discovery in Chemical Databases*, in "8th European Conference on Principles and Practice of Knowledge Discovery in

Databases - PKDD 2004, Pisa, Italy", J.-F. BOULICAUT, F. ESPOSITO, F. GIANNOTTI, D. PEDRESCHI (editors)., Lecture Notes in Artificial Intelligence, vol. 3202, Springer Verlag, September 2004, p. 39–51.

[12] S. BERASALUCE, C. LAURENÇO, A. NAPOLI, G. NIEL. *An Experiment on Mining Chemical Reaction Databases*, in "Modelling, Computation and Optimization in Information Systems and Management Sciences - MCO'04, Metz, France", L. T. H. AN, P. D. TAO (editors)., Hermes Science Publishing, London, Le Thi Hoai An and Pham Dinh Tao, July 2004, p. 535–542, http://www.loria.fr/publications/2004/A04-R-084/A04-R-084.ps.

[13] M. CADOT, J. DI MARTINO, A. NAPOLI. *Réduction d'un jeu de règles d'association par des méta-règles issues de la logique de "sens commun"*, in "Extraction et gestion des connaissances (EGC'2004), Clermont-Ferrand, France", G. HÉBRAIL, L. LEBART, J.-M. PETIT (editors)., (Poster), RNTI, Cépaduès Éditions Toulouse, 2004, 353.

[14] M. CADOT, A. NAPOLI. *Règles d'association et codage flou des données*, in "Onzièmes Rencontres de la Société Francophone de Classification (SFC-04), Bordeaux, France", M. CHAVENT, O. DORDAN, C. LACOMBLEZ, M. LANGLAIS, B. PATOUILLE (editors)., Institut de mathématiques de Bordeaux, 2004, p. 130–133.

[15] H. CHERFI, D. JANETZKO, A. NAPOLI, Y. TOUSSAINT. *Sélection de règles d'association par un modèle de connaissances pour la fouille de textes*, in "Conférence d'Apprentissage - CAp 2004, Montpellier, France", M. LIQUIÈRE, M. SEBBAN (editors)., Presses Universitaires de Grenoble, June 2004, p. 191–206, http://www.loria.fr/publications/2004/A04-R-074/A04-R-074.ps.

[16] M.-D. DEVIGNES, M. SMAÏL. *Integration of biological data from web resources: management of multiple answers through metadata retrieval*, in "12th International Conference on Intelligent Systems for Molecular Biology - 3rd European Conference on Computational Biology - ISMB-ECCB 2004, Glasgow, Scotland, United Kingdom", August 2004, http://www.loria.fr/publications/2004/A04-R-059/A04-R-059.ps.

[17] S. HERGALANT, B. AIGLE, B. DECARIS, J.-F. MARI, P. LEBLOND. *Classification non supervisée par HMM de sites de fixation de facteurs de transcription chez les bactéries*, in "5èmes Journées Ouvertes: Biologie, Informatique et Mathématiques - JOBIM'04, Montréal, Canada", June 2004, http://www.loria.fr/publications/2004/A04-R-119/A04-R-119.ps.

[18] D. JANETZKO, H. CHERFI, R. KENNKE, A. NAPOLI, Y. TOUSSAINT. *Knowledge-based Selection of Association Rules for Text Mining*, in "16h European Conference on Artificial Intelligence - ECAI'04, Valencia, Spain", R. L. DE MÀNTARAS, L. SAITTA (editors)., IOS Press, August 2004, p. 485–489, http://www.loria.fr/publications/2004/A04-R-105/A04-R-105.ps.

[19] F. LE BER, J.-F. MARI, M. BENOÎT, C. MIGNOLET, C. SCHOTT. *CarrotAge, a software for mining land-use data*, in "Fourth International Workshop on Environmental Applications of Machine Learning - EAML'2004, Bled, Slovenia", September 2004, http://www.loria.fr/publications/2004/A04-R-231/A04-R-231.ps.

[20] J. LIEBER, M. D'AQUIN, S. BRACHAIS, A. NAPOLI. *Une étude comparative de quelques travaux sur l'acquisition de connaissances d'adaptation pour le raisonnement à partir de cas*, in "12ème Atelier de Raisonnement à Partir de Cas - RàPC'04, Université Paris Nord, Villetaneuse, France", R. KANAWATI, S. SALOTTI, F. ZEHRAOUI (editors)., March 2004, p. 53–60, http://www.loria.fr/publications/2004/A04-R-

041/A04-R-041.ps.

[21] S. MAUMUS, A. NAPOLI, C. SASS, E. ALBUISSON, S. VISVIKIS. *A new approach to detect Interactions Involving Lipid Genes by Combining Data Mining and Statistics in the STANISLAS Cohort*, in "Biologie Prospective - Santorini Conference, Santorini, Greece", Poster, October 2004.

[22] L. SZATHMARY, A. NAPOLI. *Knowledge organisation and information retrieval based on Galois lattices*, in "Modelling, Computation and Optimization in Information Systems and Management Sciences - MCO '04, Metz, France", L. T. H. AN, P. D. TAO (editors)., Hermes Science Publishing, July 2004, p. 611–618.

[23] L. SZATHMARY, A. NAPOLI. *Knowledge organisation and information retrieval using Galois lattices*, in "Workshop on Knowledge Management and Organizational Memories - ECAI 2004 (16th European Conference on Artificial Intelligence), Valencia, Spain", R. DIENG-KUNTZ, N. MATTA (editors)., August 2004, p. 73–78, http://www.loria.fr/publications/2004/A04-R-120/A04-R-120.ps.

[24] L. SZATHMARY, A. NAPOLI. *Knowledge organisation and information retrieval with Galois lattices*, in "14th International Conference on Engineering Knowledge in the Age of the Semantic Web (EKAW 2004), Whittlebury Hall, UK", E. MOTTA, N. SHADBOLT, A. STUTT, N. GIBBINS (editors)., Lecture notes in Computer Science, poster, vol. 3257, Springer, October 2004, p. 511–512.

[25] L. SZATHMARY, A. NAPOLI. *Les treillis de Galois pour l'organisation et la gestion des connaissances*, in "11èmes Rencontres de la Société Francophone de Classification - SFC '04, Bordeaux, France", M. CHAVENT, O. DORDAN, C. LACOMBLEZ, M. LANGLAIS, B. PATOUILLE (editors)., September 2004, p. 298–301, http://www.loria.fr/publications/2004/A04-R-121/A04-R-121.ps.

[26] L. SZATHMARY, A. NAPOLI. *Les treillis de Galois pour l'organisation et la gestion des connaissances*, in "11èmes Rencontres de la Société Francophone de Classification - SFC '04, Bordeaux, France", M. CHAVENT, M. LANGLAIS (editors)., Revue des Nouvelles Technologies de l'Information (RNTI), vol. C, n^{o} 1, Cépaduès Éditions, September 2004, p. 153–164.

[27] M. D'AQUIN, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Decision Support and Knowledge Management in Oncology using Hierarchical Classification*, in "Proceedings of the Symposium on Computerized Guidelines and Protocols - CGP-2004, Prague, Czech Republic", K. KAISER, S. MIKSCH, S. W. TU (editors)., Studies in Health Technology and Informatics, vol. 101, IOS Press, April 2004, p. 16–30, http://www.loria.fr/publications/2004/A04-R-048/A04-R-048.ps.

[28] M. D'AQUIN, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Kasimir: gestion de connaissances décisionnelles en cancérologie*, in "Modélisation et pilotage des systèmes de Connaissances et de Compétences dans les Entreprises Industrielles - C2EI'04, Nancy, France", J. R. EMMANUEL CAILLAUD (editor)., December 2004, http://www.loria.fr/publications/2004/A04-R-512/A04-R-512.ps.

[29] M. D'AQUIN, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Kasimir: portail sémantique pour la gestion des connaissances en cancérologie*, in "Second séminaire francophone du Web Sémantique Médical - WSM 2004, Rouen, France", March 2004.

[30] M. D'AQUIN, S. BRACHAIS, J. LIEBER, A. NAPOLI. *Vers une acquisition automatique de connaissances d'adaptation par examen de la base de cas — une approche fondée sur des techniques d'extraction de*

connaissances dans des bases de données, in "12ème Atelier de Raisonnement à Partir de Cas - RàPC'04, Universite' Paris Nord, Villetaneuse, France", R. KANAWATI, S. SALOTTI, F. ZEHRAOUI (editors)., March 2004, p. 41–52, http://www.loria.fr/publications/2004/A04-R-042/A04-R-042.ps.

[31] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Représentation de points de vue pour le raisonnement à partir de cas*, in "Langages et Modèles à Objets - LMO'04, Lille, France", Revue des Sciences et Technologies de l'Information, RSTI - série L'Objet, vol. 10, n^{o} 2-3, March 2004, p. 245–258.

[32] M. D'AQUIN, J. LIEBER, A. NAPOLI. *Étude de quelques logiques de descriptions floues et de formalismes apparentés*, in "Rencontres francophones sur la logique floue et ses applications - LFA'04, Nantes, France", J. MONTMAIN (editor)., Cépadues-Éditions, November 2004, p. 255–262, http://www.loria.fr/publications/2004/A04-R-469/A04-R-469.ps.

Internal Reports

[33] R. AL HULOU, A. NAPOLI, E. NAUER. *A semantic similarity measure for content-based classification of documents*, Rapport de recherche, December 2004.

[34] R. BENDAOUD. *Fouille de données textuelles complexes*, Stage de DEA, LORIA - INRIA, June 2004, http://www.loria.fr/publications/2004/A04-R-568/A04-R-568.ps.

[35] S. BERASALUCE, G. NIEL, A. NAPOLI, C. LAURENÇO. *Data mining in reaction databases: extraction of knowledge on chemical functionality transformations*, Rapport de recherche, April 2004, http://www.loria.fr/publications/2004/A04-R-049/A04-R-049.ps.

[36] N. MESSAI. *Treillis de Galois et ontologies de domaine pour la classification et la recherche de sources de données génomiques*, Stage de DEA, June 2004, http://www.loria.fr/publications/2004/A04-R-541/A04-R-541.ps.

[37] S. OSMAN. *Réalisation d'un annuaire de sources de données génomiques en vue de la collecte et de l'intégration de données sur le Web*, Rapport de master professionnel Sciences et Techniques, mention Informatique, spécialité Bio-informatique, Stage de DESS, Universités de Bordeaux, September 2004, http://www.loria.fr/publications/2004/A04-R-545/A04-R-545.ps.

Bibliography in notes

[38] M. Y. GALPERIN. *The Molecular Biology Database Collection: 2004 update*, in "Nucleic Acids Research", vol. 32, January 2004.