Elements on KDDK: Knowledge Discovery guided by Domain Knowledge

Amedeo Napoli and the Orpailleur team

LORIA — UMR 7503
Bâtiment B, B.P. 239
F-54506 Vandœuvre-lès-Nancy cedex, France

E-mail: Amedeo.Napoli@loria.fr

Abstract. In this paper, we present and give details on the research work carried out in the Orpailleur team at LORIA, showing multiple and combined aspects of knowledge discovery and knowledge processing. The classical knowledge discovery in databases process (KDD) consists in processing a huge volume of data for extracting significant and reusable knowledge units. From a knowledge representation perspective, the KDD process may take advantage of domain knowledge embedded in ontologies relative to the domain of data, leading to the notion of KDDK, i.e. knowledge discovery (from complex data) guided by domain knowledge. The KDDK process is based on multiple forms of classification tasks, e.g. for modeling, representing, reasoning, and discovering. Various applications are introduced and detailed, showing how the notion of KDDK is instantiated. At the end of the paper, an architecture of an integrated KDDK system is proposed and discussed.

1 Introduction

In this presentation, we introduce and give details on the research work carried out in the Orpailleur team at LORIA. This is a collective research work showing multiple aspects of knowledge discovery and processing. The "orpailleur" denotes in French a person who is searching for gold in the rivers. Indeed, *knowledge discovery in databases* can be likened to the process of searching for gold in the rivers: the gold nuggets correspond to knowledge units and the rivers correspond to databases. The knowledge discovery in databases process –hereafter KDD– consists in processing a huge volume of data in order to extract knowledge units that are significant and reusable. The KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the *analyst*. The analyst selects and interprets a subset of the units for building "models" that are further considered as knowledge units with a certain plausibility.

The KDD process is performed within a KDD system including databases, data mining modules, and interfaces for interactions, e.g. editing and visualization. In this presentation, we want to emphasize the role of knowledge in the KDD

process: the KDD process may take advantage of domain knowledge embedded within an *ontology* relative to the domain of data. This leads to the notion of knowledge discovery from complex data guided by domain knowledge, or KDDK. In KDDK, the knowledge units extracted by the KDD system have still a life after the interpretation step: they must be represented in an adequate knowledge representation formalism for being integrated within an ontology and reused for problem-solving needs. In this way, the results of the KDD process may be reused for enlarging existing ontologies. The KDDK process shows that knowledge representation and KDD are two complementary processes: *no knowledge discovery without (prior) knowledge on the domain of data!*

The KDDK process is based on the core idea of *classification*. Classification is a polymorphic process involved in various tasks, e.g. modeling, mining, representing, and reasoning (see also [40, 8, 45]). Accordingly, a knowledge-based system may be designed, fed up by the KDDK process, and used for problem-solving in application domains. For the Orpailleur team, these applications domains are mainly agronomy, astronomy, biology, chemistry, and medicine. A special mention has to be made for Semantic Web activities, involving in particular text mining, content-based document mining, and intelligent information retrieval (see for example [15, 6, 39]).

The KDD process is based on *data mining methods* that are either symbolic or numerical [18, 19, 13]. In the Orpailleur team, KDD is both from symbolic and numerical points of view:

- Symbolic methods are mainly based on lattice-based classification (concept lattice design or formal concept analysis [17]), frequent itemsets search, and association rule extraction [35].
- Numerical methods are mainly based on Hidden Markov Models of order 1 and 2 (initially designed for pattern recognition) [30].

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

2 Methods and systems for KDD

2.1 Lattice design, itemset search and association rule extraction

Classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), and a binary correspondence

between the two classes, indicating for each individual-property pair whether the property applies to the individual or not [3, 17, 6]. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Lattice-based classification relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting (from a binary database) a set of concepts organized within a hierarchy (i.e. a partial ordering). The extraction of frequent itemsets, i.e. sets of properties or features of data occurring together with a certain frequency, and of association rules emphasizing correlations between sets of properties with a given confidence, are related activities.

The search for frequent itemsets and association rule extraction are well-known symbolic data mining methods. These processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications.

Accordingly, the CORON platform is currently developed in the team [42]. The platform is composed of three main modules: (i) CORON-base, (ii) ASSRULEX, (iii) pre-processing and post-processing modules. The CORON-base module is aimed at extracting different kinds of itemsets, e.g. frequent itemsets, frequent closed itemsets, minimal generators, etc. The module contains a collection of important data mining algorithms, such as Apriori, Close, Apriori-Close, Pascal, Titanic, Charm, Eclat, together with adapted algorithms such as Pascal$^+$, Zart, RMS Carpathia, Eclat-Z. This large collection of (efficient) algorithms is one of the main characteristics of the CORON platform. Knowing that each of the algorithms has advantages and disadvantages with respect to the form of the data to be mined, and since there is no universal algorithm for processing any arbitrary dataset, the CORON-base module offers to the user the choice of the algorithm that is the best suited for his needs.

The second module of the system, ASSRULEX (Association Rule eXtractor) generates different sets of association rules, such as informative rules, generic basis, and informative basis.

For supporting the whole life-cycle of a data mining task, the CORON platform proposes 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.

The CORON platform is developed entirely in Java, allowing portability. The system is operational, and has been tested within several research projects within the team [10, 31] (see figure 1).

2.2 Stochastic methods for KDD

Among numerical methods for data mining, the Orpailleur team is mainly interested in stochastic models based on second-order Hidden Markov Models (HMM2)

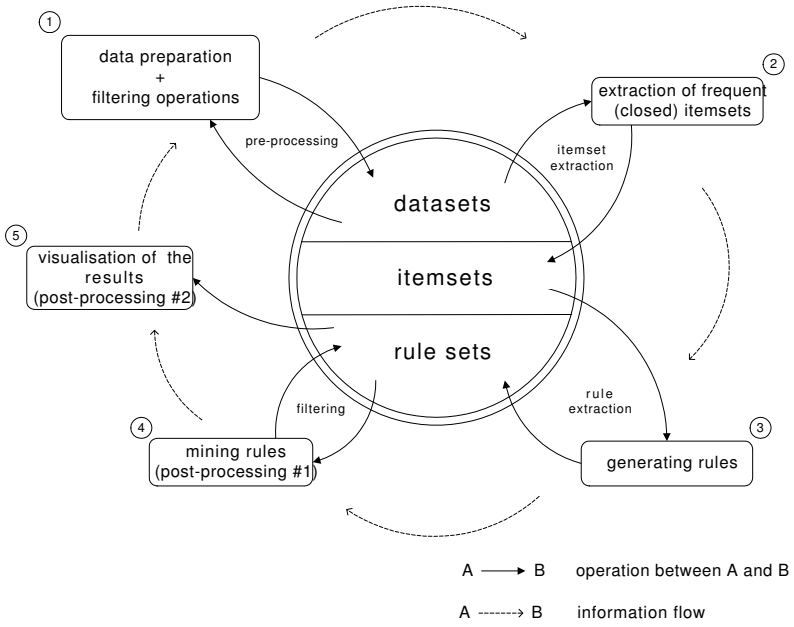

Fig. 1. The life cycle of KDD within Coron.

for mining temporal and spatial data. Hidden Markov Models have good capabilities to locate stationary segments (as shown in research work on speech recognition [29]). These models map sequences of data into a Markov chain in which transitions between states depend on the $\mathtt{n}$ previous states according to the order of the model ($\mathtt{n} = 2$ for HMM2). Actually, a second-order Hidden Markov model is defined as follows: (i) a set $\mathtt{S} = (\mathtt{s}_1, \ldots \mathtt{s}_N)$ of $\mathtt{N}$ states, (ii) a 3 dimensional matrix on $\mathtt{S}^3$ with $\mathtt{a}_{ijk} = \mathtt{Prob}(\mathtt{q}_t = \mathtt{s}_k / \mathtt{q}_{t-1} = \mathtt{s}_j, \mathtt{q}_{t-2} = \mathtt{s}_i)$, where $\mathtt{q}_t$ denotes the state at time $\mathtt{t}$ and $\sum_{k=1}^{N} \mathtt{a}_{ijk} = 1, \forall (\mathtt{i}, \mathtt{j}) \in [1, \mathtt{N}] \times [1, \mathtt{N}]$, (iii) a set of $\mathtt{N}$ discrete distributions: $\mathtt{b}_i(.)$ denotes for $\mathtt{i}$ the distribution of observations associated to the state $\mathtt{s}_i$. This distribution may be parametric, non parametric, or even given by another Hidden Markov Model. The process of mining databases with HMM2 may be considered as an unsupervised classification process, where domain knowledge is modeled as a sequence of states resulting from a sequence of observations.

The CAROTTAGE system[1] is developed in the team for mining numerical spatio-temporal data using HMM2. The purpose of the system is to to analyze spatio-temporal data by building a partition of homogeneous classes in temporal

[1] CAROTTAGE is a free software developed in the Orpailleur team, with a GPL license since 2002, see `http://www.loria.fr/~jfmari/App/`.

and spatial dimensions, with a view on the transitions between the classes. The system takes as input an array of discrete data, where the rows represent the spatial sites and the columns the time slots, and 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.

The CAROTTAGE system has been involved for data mining purposes in two main application domains, namely biology and agronomy. In collaboration with biologists, genome segmentation and interpretation have been investigated [20, 14]. In collaboration with agronomists, spatial and temporal land-use data have been mined for extracting and understanding crop successions, i.e. the way how crops are carried out during a given period of time [25, 30]. In these two applications, the effort has focused on two main points, with respect to the questions of the biologists and of the agronomists: (i) the elaboration of a mining process for extracting dependencies in temporal and spatial data involving an unsupervised classification process based on HMM2, (ii) the specification of associated and adequate visualization tools giving a synthetic view of the extraction process results to the experts in charge of interpreting the extracted classes and/or of specifying new experiment directions.

3 Research directions for KDDK

The principle summarizing KDDK can be read as follows: going "from complex data units to complex knowledge units guided by domain knowledge" (KDDK) or "knowledge with/for knowledge". This principle is implemented in the present and future work of the Orpailleur team and is discussed below, along research activities such as graph mining, spatio-temporal data mining, text mining and Semantic Web, knowledge discovery in life sciences, combining symbolic and numerical data mining methods for hybrid mining, and finally mining a knowledge base, a kind of "meta-knowledge discovery process". All these research activities share the fact that the mining process is guided and enhanced by domain knowledge (similar ideas are also discussed in [9, 45]).

3.1 Some extensions of standard mining methods

Lattice-based classification, formal concept analysis, itemset search and association rule extraction, are suitable paradigms for symbolic KDDK, that may be used for real-sized applications [44]. Global improvements may be carried on the ease of using of the data mining methods, on the efficiency of the methods [24], and on adaptability, i.e. the ability to fit evolving situations with respect to the constraints that may be associated with the KDDK process. Accordingly, a first research line is the extension of symbolic methods to complex data, e.g. objects with multi-valued attributes, relations, and graphs [23]. A second research line is the search for rare itemsets, i.e. itemsets whose frequency is under a certain

threshold [41]. This kind of search is of a valuable interest for understanding rare diseases or unexpected events.

The mining of chemical chemical reaction databases can be used for illustrating the first point. This task is important for at least two reasons: (i) the first reason is the challenge represented by this task regarding KDDK to be set on, (ii) the second reason lies in the industrial needs that can be met whenever substantial results are obtained. Chemical reactions are complex data, that may be modeled as undirected labeled graphs. They are the main elements on which synthesis in organic chemistry relies, knowing that synthesis —and accordingly chemical reaction databases— are of first importance in chemistry, but also in biology, drug design, and pharmacology. From a problem-solving point of view, synthesis in organic chemistry must be considered at two main levels of abstraction: a strategic level where general synthesis methods are involved –a kind of meta-knowledge– and a tactic level where specific chemical reactions are applied. An objective for improving computer-based synthesis in organic chemistry is aimed at discovering general synthesis methods from currently available chemical reaction databases for designing generic and reusable synthesis plans.

A preliminary research work has been carried on in the Orpailleur team [4], based on frequent levelwise itemset search and association rule extraction, and applied to standard chemical reaction databases. This work has given substantial results for the expert chemists. At the moment, for extending this first work, a graph-mining process is used for extracting knowledge from chemical reaction databases, directly from the molecular structures and the reactions themselves, This research work is currently under development, in collaboration with chemists, and is in accordance with needs of chemical industry [36].

Temporal and spatial data are complex data to be mined because of their internal structure, that can be considered as multi-dimensional. Indeed, spatial data may involve two or three dimensions for determining a region and complex relations as well for describing the relative positions of regions between each others (as in the RCC-8 theory for example [26]). Temporal data may present a linear but also a two-dimensional aspect, when time intervals are taken into account and have to be analyzed (using Allen relations for example). In this way, mining temporal or spatial data are tasks related to KDDK. Spatial and temporal data may be analyzed with numerical methods such as Hidden Markov Models, but also with symbolic methods, such as levelwise search for frequent sequential or spatial patterns.

In the medical domain, the study of chronic diseases is a good example of KDDK process on spatio-temporal data. An experiment for characterizing the patient pathway using the extraction of frequent patterns, sequential and not sequential, from the data of the PMSI[2] system associated with the "Lorraine Region" is currently under investigation. Details on this work are given in this volume [22].

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

3.2 KDDK, text mining and Semantic Web

The objective of a text mining process is to extract new and useful knowledge units in a large set of texts [21, 8, 7]. The text mining process shows some specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making the mining process more complex. To avoid information dispersion, a text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary, and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods. In addition, from a KDDK perspective, the text mining process is aimed at extracting new knowledge units from texts with the help of background knowledge. The interpretation of a text relies on knowledge units shared by the authors and the readers. A part of these knowledge units is expressed in the texts and may be extracted by the text mining process. Another part of these knowledge units, background knowledge, is not explicitly expressed in the text and is useful to relate notions present in a text, to guide and to help the text mining process. Background knowledge is encoded in a knowledge base associated to the text mining process. Text mining is especially useful in the context of semantic Web, for manipulating textual documents by their content.

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

Semantic Web constitutes a good platform for experimenting ideas on knowledge discovery –especially text mining–, knowledge representation and reasoning. In particular, the knowledge representation language associated with the Semantic Web is the OWL language, based on description logics (or DLs, see [2]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation that is a partial ordering. The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Concept classification is used for inserting a new concept at the right location in the concept hierarchy, searching for its most specific subsumers and its most general subsumees. Individual classification is used for recognizing the concepts an individual may be an instance of. Furthermore, classification-based reasoning may be extended into case-based reasoning (CBR), that relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to

the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem. When there is enough interest, the target problem and its solution may be memorized in the case base to be reused. In the context of a concept hierarchy, retrieval and adaptation may be both based on classification (and "adaptation-guided retrieval" [16]).

In the framework of Semantic Web, the mining of textual documents on the Web, or "Web document mining" [5], can be considered from two main points of view: (i) mining the content of documents, involving text mining, (ii) mining the internal and external –hypertext links– structure of pages, involving information extraction. Web document mining is a major technique for the semi-automatic design of real-scale ontologies, the backbone of Semantic Web. In turn, ontologies are used for annotating the documents, enhancing document retrieval and document mining. In this way, Web document mining improves annotation, retrieval, and the understandability of documents, with respect to their structure and their content. The extracted knowledge units can then be used for completing domain ontologies, that, in turn, guide text mining, and so on.

A research carried on in the team aims at understanding the structure of documents for analyzing and for improving text mining. The design of a system for extracting information units –that have to be turned into knowledge units after interpretation– from Web pages involves a wrapper-based machine learning algorithm combined with a classification-based reasoning process, taking advantage of a domain ontology implemented within the Web Ontology Language (OWL). The elements returned by the process are used as "semantic annotations" for understanding and manipulating the documents with respect to their structure and content [43]. The application domain of this research work is the study of research themes in the European Research Community. This study supports the analysis of research themes and detection of research directions.

3.3 KDDK for life science

Knowledge discovery is gaining more and more interest and importance in life sciences for mining either homogeneous databases (DBs) such as protein sequences or structures, heterogeneous DBs for discovering interactions between genes and environment, or between genetic and phenotypic data, especially for public health and pharmacogenomics domains. The latter case appears to be one main challenge in knowledge discovery in biology and involves knowledge discovery from complex data and thus KDDK. The interactions between researchers in biology and researchers in computer science improve not only knowledge about systems in biology, but knowledge about computer science as well. Solving problems for biologists using KDDK methods may involve the design of specific modules that, in turn, leads to adaptations of the KDDK process, especially in the preparation of data and in the interpretation of the extracted units.

A research work carried on in the team is in concern with the search and the access to relevant biological sources (including biological DBs) satisfying a set of given constraints, expressed with respect to concepts lying in a domain

ontology –as in the BioRegistry repository [38]. The sources may be described in terms of these concepts, yielding a formal context, from which a concept lattice can be built [32]. Given a specific query, a lattice-based information retrieval process is set on. The classification of the query in the lattice returns a ranked list of relevant sources, according to the characteristics of the sources with respect to the characteristics of the query (see also [33], this volume). The next step is to generalize the approach, and to use a "fuzzy concept lattice" and "fuzzy formal concept analysis" (see for example [37]). Moreover, studies hold on complex question answering methods taking into account fuzzy concept lattices, nested queries (intersection, union, and complement), analogical queries, and composition of answers elements. These techniques are still under study.

Another challenge is to extract knowledge from heterogeneous DBs for understanding interactions between clinical, genetic and therapeutic data. For example, a given genotype, i.e. a set of selected gene versions, may explain adverse clinical reactions (e.g. hyperthermy, toxic reaction...) to a given therapeutic treatment. This requires first the integration of both genomic and clinical data into a data warehouse on which KDDK methods have to be applied. This research work is connected with Semantic Web purposes, and in particular with the following elements: (i) data preparation and extracted units interpretation based on domain ontologies, (ii) knowledge edition for building and enriching domain ontologies, (iii) knowledge management for access to knowledge units, querying and reasoning (for problem-solving).

3.4 Combining symbolic and numerical methods for KDDK

The combination of symbolic and numerical data mining methods relies on HMM2 and on symbolic methods, e.g. for reasoning such as CBR or for symbolic KDD such as concept lattice design. A challenge is to set on a methodology for hybrid KDDK, coupling HMM2 and symbolic methods, that can be adapted and reused as a general KDDK method on various data, leading to a multi-functional and multi-purpose KDDK system.

Following this line, HMM2 have proved to be a valuable tool for extracting knowledge from complex numerical data, e.g. spatio-temporal data. However, some operations remain very difficult to be carried out and could be eased using symbolic methods: (i) the modeling of the HMM2 process for a set of given data, (ii) the interpretation of units extracted by HMM2, (iii) the organization and the visualization of the extracted units for further reuse, e.g. as knowledge units in a knowledge-based system. A proposition is to combine HMM2 with symbolic methods, such as case-based reasoning and concept lattices, for helping the modeling and interpretation process.

Case-based reasoning seems to be especially interesting since researchers in an application domain often use their own knowledge or knowledge resulting from first experiments to improve steps within the data mining process, e.g. modeling and interpretation. In this way, the elements of the cases within the case-based reasoner can be composed of knowledge units about parameters of the HMM2, and as well of knowledge units on the design –modeling, data preparation–, and the

interpretation –relying on ontological knowledge– of the HMM2. In addition, CBR can be of great interest for recording mining strategies that can be adapted and reused in similar situations. Indeed, a study on CBR for guiding mining scenarios in a given situation –with retrieval and adaptation of a similar situation– has not yet been carried on and should give substantial results. More generally, HMM2-based data mining process may take advantage of being coupled with CBR, that can be used at a strategic level for guiding the HMM2-based data mining process.

For their part, concept lattices can be used to organize and to visualize the results of the HMM2-based data mining process. The objects resulting of the application of the HMM2 process can be characterized by a set of properties. For example, in a spatio-temporal framework, space regions may be considered as objects and characteristics of the region at a given time can be considered as properties, yielding a kind of formal context. In addition, itemsets and association rules may also be extracted from such a context, offering an easy way of interpreting results of the HMM2 process.

For concluding, the analysis of complex data in biology also calls for the coupling of symbolic and numerical data mining methods. There are complex data on which HMM2 show a good behavior, for recognizing and extracting regular structures. Such complex data hold on interactions between processes or agents, such as data from transcriptomic biochips –DNA chips or microarrays– experiments (used for extracting knowledge on interactions between plants and microorganisms). Still, an important objective of this kind of study is to investigate and to understand more deeply the modeling of biological systems, at symbolic and numerical levels.

3.5 Meta-knowledge discovery of mining knowledge bases

The main tasks of the KASIMIR system are decision support and knowledge management for the treatment of cancer. The system is developed within a multidisciplinary research project in which participate researchers from different community (computer science, ergonomics, and oncology). For a given cancer localization, a treatment is based on a protocol similar to a medical guideline. 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 said to be *out-of-the-protocol*, i.e. either the protocol does not provide a treatment for this medical case, or the proposed solution raises some difficulties, e.g. contraindication, treatment impossibility, etc. For such an out-of-the-protocol case, oncologists try to *adapt* the protocol. In turn, 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.

In knowledge-intensive CBR, the reuse of cases is generally based on adaptation, the goal of which is to solve the target problem by adapting the solution of a source case. The adaptation process is based on adaptation knowledge that –for the main part– is domain-dependent, and thus needs to be acquired for a

new application of CBR. Adaptation knowledge plays a key issue in applications, e.g. in knowledge-intensive case-based reasoning systems [1].

In parallel, the Semantic Web technology relies on the availability of large amount of knowledge in various forms [15, 39]. The acquisition of ontologies is one of the important issues that is widely explored in the Semantic Web community. Moreover, the acquisition of decision and adaptation knowledge for the Semantic Web has not been so deeply explored, though this kind of knowledge can be useful in numerous situations. For example, given a decision protocol and an adaptation knowledge base, the KASIMIR system can be used to apply and/or to adapt the protocol to specific medical situations.

The goal of *adaptation knowledge acquisition* (AKA) is to mine a case base, to extract adaptation knowledge units, and to make these units operational. Until now, the research work on CBR in the Orpailleur team has mainly focused on the design of algorithms and knowledge representation formalisms for implementing the adaptation process in a CBR system. A next step is to investigate the AKA process, a research topic that has still not received so much in the CBR community. A parallel research topic is to apply AKA to the extraction of decision knowledge units. Indeed, adaptation knowledge is closely related with decision theory, e.g. the Wald pessimistic criterion is frequently applied when pieces of information about a patient are missing [11].

Accordingly, the objective of the research work on AKA is to study how KDD techniques can be used for feeding a knowledge server embedded in a semantic portal –as the KASIMIR semantic portal [11]– and thus to instantiate the KDDK process. In the KASIMIR semantic portal, OWL-based formalisms for representing medical ontologies, decision protocols (the case base), and adaptation knowledge, are designed. Web services associated to the CBR process are developed. Several protocols are implemented, with a few of them including adaptation knowledge.

Practically, AKA can be considered from two main points of view. AKA from experts is based on 'manual" analysis of documents related to current problems. The AKA from expert process leads to the elaboration of *adaptation rules*, depending on formal parameters and associated with explanations. The adaptation rules are human-understandable –thanks to explanations– but they need additional knowledge for instantiating the parameters and being applied (more on AKA from experts is given in [27, 28, 34]).

Semi-automatic AKA is based on the principles of KDD, and involves data preparation, data mining, and interpretation of the extracted units, under the control of an analyst. The input of the AKA process is a set of adaptations –thus elements at the knowledge level– and the output is a set of adaptation rules. Such an adaptation rule is an operational association rule, that lack explanations. Mixed AKA combines AKA from experts and semi-automatic AKA for supplying operational and human-understandable adaptation knowledge.

In the current experiments within the KASIMIR system, semi-automatic AKA is based on frequent itemset search. A system for AKA, named CABAMAKA–case base mining for AKA, is currently under development within the KASIMIR system and relies on semi-automatic AKA [10, 12]. The CABAMAKA system analyzes a

simple representation of the variations $\Delta\mathtt{u}$ between units of knowledge $\mathtt{u}_1$ and $\mathtt{u}_2$, where $\Delta\mathtt{u}$ encodes the substitutions transforming $\mathtt{u}_1$ into $\mathtt{u}_2$. The variations are represented in an expressive DL-based formalism, allowing a high-level expression of the extracted adaptation rules. Beyond CBR, such a research work can be useful for ontology alignment: an alignment expresses a correspondence between the elements of two ontologies, but it could also express the variations between corresponding elements, within a rich representation formalism for the variations.

4 Towards an integrated KDDK system

From a global point of view, the research objectives for KDDK can be summarized as follows:

- A methodology for a "knowledge discovery from complex data guided by domain knowledge process" (KDDK), i.e. a process leading from complex data units to complex knowledge units taking advantage of domain knowledge, at each step of the knowledge discovery process.
- A combination of symbolic and numerical data mining methods for setting up a complete and hybrid mining methodology to be applied on various types of data.
- An implementation of the "knowledge discovery from complex data guided by domain knowledge process" within an operational system, to be used on a large set of data types, e.g. textual documents, genomic data, spatio-temporal data, graphs, and even on sets of knowledge units (a kind of meta-knowledge mining), i.e. mining a knowledge base instead of a database.
- Accordingly, the design of a KDDK system, based on the above principles, and involved in application domains such as astronomy, agronomy, biology, chemistry, medicine, for decision support and problem-solving.

From a middle-term perspective, a system for KDDK can be considered as a "decentralized system" the architecture of which is described hereafter.

- One or several ontologies (knowledge bases) include knowledge from different domains with different points of view, and as well, a case base. A set of services are related through a semantic portal, for knowledge editing, navigating, and visualizing the ontologies.
- An inference engine provides, in association with the knowledge bases, a collection of inference rules for problem-solving purposes, among which subsumption, classification (lattice-based classification, clustering), case-based reasoning. Reasoning services are present for handling concrete datatypes such as strings or numbers (and possibly, for controlling procedural or functional reasoning modes if-needed).
- A set of heterogeneous databases holding on a domain to be mined for providing knowledge units enriching domain ontologies.
- A platform for KDDK proposes a collection of data mining modules –such as the CORON platform– and a set of services for data preparation and extracted unit interpretation.

Moreover, the system has to provide channels for allowing communications with human agents, such as experts and end-users. The resulting KDDK system architecture has to be reusable in any application domain. Accordingly, the integration of such a KDDK system in the framework of the semantic Web can be seen as follows. The data sources, i.e. databases, sets of documents, are explored, navigated, and queried, under the supervision of an analyst, thanks to a KDDK process guided by knowledge bases of the domain. The data are prepared and manipulated by the KDDK process, while the knowledge units are validated by the analyst, and then manipulated by the inference engine.

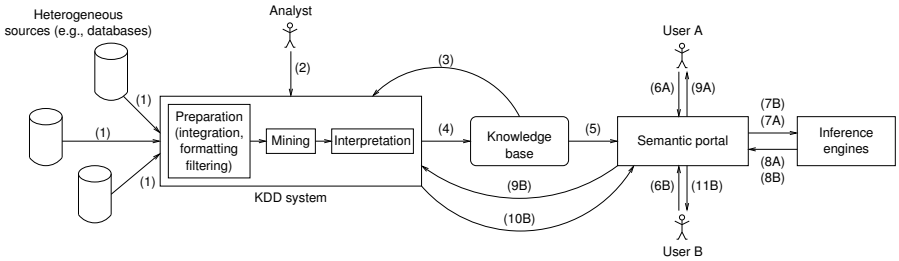

Fig. 2. An architecture for a system aimed at "knowledge discovery (from complex data) guided by domain knowledge process (KDDK)". The classical KDD process can be read from left to right, while, by contrast, the KDDK system can be read from right to left.

The figure 2 presents the architecture proposal for a KDDK system, in which different scenarios can be made operational. Heterogeneous sources (e.g. databases) feed the KDD system (1), under the supervision of an analyst (2), using available domain knowledge (3). The KDD system returns new knowledge units for extending and enriching a knowledge base (4), that may be queried through a semantic portal (5) by distant geographically distributed users (users A and B). The users A and B query the portal (6A, 6B), that in turn may use the services of a knowledge base and the associated inference engine (7A, 7B). When the available knowledge provides, with the help of the inference engine, an answer to the request (8A), this answer is transmitted to the user (9A). Otherwise (8B), the request is transferred in a filtering module used by the KDD system (9B) for mining the available data, trying to extract information related to the request. The resulting extracted knowledge units relying on this filter (10B) may provide an answer to the user (11B).

5 Conclusion

In this paper, we have presented the research work carried out in the Orpailleur team at LORIA. Multiple and combined aspects of knowledge discovery and

knowledge processing have been introduced and discussed: symbolic KDD methods such as lattice-based classification itemset search, and association rule extraction, and numeric methods such as HMM2. Next, the KDD process has been considered from a knowledge representation perspective, explaining how and why the KDD process may take advantage of domain knowledge embedded in ontologies relative to the domain of data. This perspective leads to the idea of KDDK, for knowledge discovery (from complex data) guided by domain knowledge. The KDDK process is based on classification tasks, for modeling, representing, reasoning, and discovering. Various instantiations of the KDDK process have been described, among which the mining of molecular graphs –for knowledge discovery in chemical reaction databases–, text mining and Semantic Web for designing and enlarging ontologies from documents, knowledge discovery in life sciences, and hybrid knowledge discovery, combining numerical and symbolic methods for data mining. An original experiment has also been introduced and discussed: meta-knowledge mining, or mining a knowledge base instead of a database. This research work has been carried out for the need of adaptation knowledge acquisition (AKA), that is a promising research domain, and that can be reused for mining various kind of strategical knowledge units, e.g. decision knowledge units. At the end of the paper, an architecture of an integrated KDDK system has been proposed and discussed.

References

1. A. Aamodt. Knowledge-Intensive Case-Based Reasoning and Sustained Learning. In L. C. Aiello, editor, *Proc. of the 9th European Conference on Artificial Intelligence (ECAI'90)*, 1990.
2. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, Cambridge, UK, 2003.
3. M. Barbut and B. Monjardet. *Ordre et classification – Algèbre et combinatoire (2 tomes)*. Hachette, Paris, 1970.
4. S. Berasaluce, C. Laurenço, A. Napoli, and G. Niel. An Experiment on Knowledge Discovery in Chemical Databases. In J.-F Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, *Knowledge Discovery in Databases: PKDD 2004, Pisa*, Lecture Notes in Artificial Intelligence 3202, pages 39–51. Springer, Berlin, 2004.
5. B. Berendt, A. Hotho, and G. Stumme. Towards Semantic Web Mining. In I. Horrocks and J. Hendler, editors, *The Semantic Web - ISWC 2002*, Lecture Notes in Artificial Intelligence 2342, pages 264–278. Springer, Berlin, 2002.
6. C. Carpineto and G. Romano. *Concept Data Analysis: Theory and Applications*. John Wiley & Sons, Chichester, UK, 2004.
7. H. Cherfi, A. Napoli, and Y. Toussaint. Towards a text mining methodology using association rules extraction. *Soft Computing*, 10(5):431–441, 2006.
8. P. Cimiano, A. Hotho, and S. Staab. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.
9. P. Cimiano, A. Hotho, G. Stumme, and J. Tane. Conceptual knowledge processing with formal concept analysis and ontologies. In P.W. Eklund, editor, *Concept*

Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, Lecture Notes in Computer Science 2961, pages 189–207. Springer, 2004.

10. M. d'Aquin, F. Badra, S. Lafrogne, J. Lieber, A. Napoli, and L. Szathmary. Adaptation knowledge discovery from a case base. In G. Brewka, S. Coradeschi, A. Perini, and P. Traverso, editors, *17h European Conference on Artificial Intelligence – ECAI'06, Riva del Garda, Italy*, pages 795–796, 2006.
11. M. d'Aquin, C. Bouthier, S. Brachais, J. Lieber, and A. Napoli. Knowledge Edition and Maintenance Tools for a Semantic Portal in Oncology. *International Journal on Human–Computer Studies*, 62(5):619–638, 2005.
12. M. d'Aquin, S. Brachais, J. Lieber, and 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 R. Kanawati, S. Salotti, and F. Zehraoui, editors, *12ième Atelier de Raisonnement à Partir de Cas - RàPC'04, Université Paris Nord, Villetaneuse, France*, pages 41–52, 2004.
13. M.H. Dunham. *Data Mining – Introductory and Advanced Topics*. Prentice Hall, Upper Saddle River, NJ, 2003.
14. Catherine Eng, Annabelle Thibessard, Sébastien Hergalant, Jean-François Mari, and Pierre 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*, 2005.
15. D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors. *Spinning the Semantic Web*. The MIT Press, Cambridge, Massachusetts, 2003.
16. B. Fuchs, J. Lieber, A. Mille, and A. Napoli. An Algorithm for Adaptation in Case-based Reasoning. In W. Horn, editor, *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin*, pages 45–49. IOS Press, Amsterdam, 2000.
17. B. Ganter and R. Wille. *Formal Concept Analysis*. Springer, Berlin, 1999.
18. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2001.
19. D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. The MIT Press, Cambridge (MA), 2001.
20. Sébastien Hergalant, Bertrand Aigle, Bernard Decaris, Jean-François Mari, and Pierre 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*, Jun 2004.
21. D. Janetzko, H. Cherfi, R. Kennke, A. Napoli, and Y. Toussaint. Knowledge-based selection of association rules for text mining. In R. Lopez de Màntaras and L. Saitta, editors, *16h European Conference on Artificial Intelligence – ECAI'04, Valencia, Spain*, pages 485–489, 2004.
22. N. Jay, F. Kohler, and A. Napoli. Using formal concept analysis for mining and interpreting patient flows within a healthcare network. In S. Ben Yahia and E. Mephu-Nguifo, editors, *Fourth International Conference on Concept Lattices and their Applications (CLA-06), Hammamet, Tunisia*. Springer, 2006.
23. S.O. Kuznetsov. Machine learning and formal concept analysis. In Peter W. Eklund, editor, *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia*, Lecture Notes in Computer Science 2961, pages 287–312. Springer, 2004.

24. S.O. Kuznetsov and S.A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Theoretical Artificial Intelligence*, 14(2/3):189–216, 2002.
25. F. Le Ber, M. Benoit, C. Schott, J.-F. Mari, and C. Mignolet. Studying crop sequences with CarrotAge, a HMM-based data mining software. *Ecological Modelling*, 191(1):170 – 185, 2006.
26. F. Le Ber and A. Napoli. Design and comparison of lattices of topological relations for spatial representation and reasoning. *Journal of Experimental & Theoretical Artificial Intelligence*, 15(3):331–371, 2003.
27. J. Lieber, P. Bey, F. Boisson, B. Bresson, P. Falzon, A. Lesur, A. Napoli, M. Rios, and C. Sauvagnac. Acquisition et modélisation de connaissances d'adaptation, une étude pour le traitement du cancer du sein. In *Actes des journées ingénierie des connaissances (IC-2001)*, pages 409–426, Grenoble, 2001.
28. J. Lieber, M. d'Aquin, P. Bey, A. Napoli, M. Rios, and C. Sauvagnac. Adaptation knowledge acquisition, a study for breast cancer treatment. In M. Dojat, E. Keravnou, and P. Barahona, editors, *9th Conference on Artificial Intelligence in Medicine in Europe2003 – AIME 2003, Protaras, Chypre*, LNCS 2780, pages 304–313. Springer, Berlin, 2003.
29. J.-F. Mari, J.-P. Haton, and A. Kriouile. Automatic Word Recognition Based on Second-Order Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 5:22 – 25, 1997.
30. J.-F. Mari and F. Le Ber. Temporal and spatial data mining with second-order hidden models. *Soft Computing*, 10(5):406–414, 2006.
31. S. Maumus, A. Napoli, L. Szathmary, and 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*, pages 169–173, 2005.
32. N. Messai, M.-D. Devignes, A. Napoli, and M. Smaïl-Tabbone. Querying a bioinformatic data sources registry with concept lattices. In F. Dau, M.-L. Mugnier, and G. Stumme, editors, *Conceptual Structures: Common Semantics for Sharing Knowledge, Proceedings of the 13th International Conference on Conceptual Structures, ICCS 2005, Kassel, Germany*, Lecture Notes in Computer Science 3596, pages 323–336, 2005.
33. N. Messai, M.-D. Devignes, A. Napoli, and M. Smaïl-Tabbone. Br-explorer: An fca-based algorithm for information retrieval. In S. Ben Yahia and E. Mephu-Nguifo, editors, *Fourth International Conference on Concept Lattices and their Applications (CLA-06), Hammamet, Tunisia*. Springer, 2006.
34. V. Mollo. *Usage des ressources, adaptation des savoirs et gestion de l'autonomie dans la décision thérapeutique*. Thèse d'Université, Conservatoire National des Arts et Métiers, 2004.
35. A. Napoli. A smooth introduction to symbolic methods for knowledge discovery. In H. Cohen and C. Lefebvre, editors, *Handbook of Categorization in Cognitive Science*, pages 913–933. Elsevier, Amsterdam, 2005.
36. F. Pennerath and A. Napoli. La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique. In G. Ritschard and C. Djeraba, editors, *Extraction et gestion des connaissances (EGC'2006), Lille*, pages 517–528. RNTI-E-6, Cépaduès-Éditions Toulouse, 2006.
37. T.T. Quan, S.C. Hui, A.C.M. Fong, and T.H. Cao. Automatic generation of ontology for scholarly semantic web. In S.A. McIlraith, D. Plexousakis, and F. Van

Harmelen, editors, *International Conference on Sematic Web, ISWC 2004, Hiroshima, Japan*, Lecture Notes in Computer Science 3298, pages 726–740. Springer, 2004.
38. M. Smaïl-Tabbone, S. Osman, N. Messai, A. Napoli, and M.-D. Devignes. Bioregistry : a structured metadata repository for bioinformatic databases. In *First International Symposium on Computational Life Science - CompLife 2005, Konstanz, Germany*, Lecture Notes in Computer Science 3695, pages 46–56, 2005.
39. S. Staab and R. Studer, editors. *Handbook on Ontologies*. Springer, Berlin, 2004.
40. G. Stumme. Formal concept analysis on its way from mathematics to computer science. In U. Priss, D. Corbett, and G. Angelova, editors, *Conceptual Structures: Integration and Interfaces, Proceedingsof the 10th International Conference on Conceptual Structures, ICCS 2002, Borovets, Bulgaria*, Lecture Notes in Artificial Intelligence 2393, pages 2–19, Berlin, 2002. Springer.
41. L. Szathmary, S. Maumus, P. Petronin, Y. Toussaint, and A. Napoli. Vers l'extraction de motifs rares. In G. Ritschard and C. Djeraba, editors, *Extraction et gestion des connaissances (EGC'2006), Lille*, pages 499–510. RNTI-E-6, Cépaduès-Éditions Toulouse, 2006.
42. L. Szathmary and A. Napoli. Coron: A framework for levelwise itemset mining algorithms. In B. Ganter, R. Godin, and E. Mephu Nguifo, editors, *Third International Conference on Formal Concept Analysis (ICFCA'05), Lens, France, Supplementary Proceedings*, pages 110–113, 2005. Supplementary Proceedings.
43. Sylvain Ténier, Amedeo Napoli, Xavier Polanco, and Yannick Toussaint. Semantic annotation of webpages. In S. Handschuh, editor, *Workshop on Knowledge Markup and Semantic Annotation – SemAnnot 2005, ISWC 2005 Workshop, Galway, Irlande*, 2005.
44. P. Valtchev, R. Missaoui, and R. Godin. Formal concept analysis for knowledge discovery and data mining: The new challenges. In Peter W. Eklund, editor, *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia*, Lecture Notes in Computer Science 2961, pages 352–371. Springer, 2004.
45. R. Wille. Mathods of conceptual knowledge processing. In R. Missaoui and J. Schmid, editors, *International Conference on Formal Concept Analysis, ICFCA 2006, Dresden, Germany*, Lecture Notes in Artificial Intelligence 3874, pages 1–29. Springer, 2006.