Managing complex knowledge in natural sciences

Noël Conruyt, David Grosser

IREMIA, Institut de REcherche en Mathématiques et Informatique Appliquées
University of La Réunion
15, av. René Cassin – 97715 Saint-Denis, Messag. Cedex 9, France
{Conruyt, Grosser}@univ-reunion.fr

Abstract. In many fields dependant upon complex observation, the structuring, depiction and treatment of knowledge can be of great complexity. For example in Systematics, the scientific discipline that investigates bio-diversity, the descriptions of specimens are often highly structured (composite objects, taxonomic attributes), noisy (erroneous or unknown data), and polymorphous (variable or imprecise data). In this paper, we present IKBS, an Iterative Knowledge Base System for dealing with such complex phenomena. The originality of this system is to implement the scientific method in biology: experimenting (learning rules from examples) and testing (identifying new individuals, improving the initial model and descriptions). This methodology is applied in the following ways in IKBS: 1 - Knowledge is acquired through a descriptive model that suits the semantic demand of experts. 2 - Knowledge is processed with an algorithm derived from C4.5 in order to take into account structured knowledge introduced in the previous descriptive model of the domain. 3 - Knowledge is refined through the use of an iterative process to evaluate the robustness of the descriptive model and descriptions. The IKBS system is presented here as a life science application facilitating the identification of coral specimens of the family *Pocilloporidæ*.

1. Introduction

In the natural sciences, data to be processed may be more complex than in other fields. In Systematics, attributes that describe organisms are numerous (> 100) compared with the number of individuals by class which is mostly not representative (< 10): the domain to describe is established deterministically (empirically) rather than probabilistically (statistically) [14]. In such domains, we must take into account diversity and incompleteness, and the exception is the only valid rule.

Learning systems intended to facilitate classification (class definition) and identification of natural organisms must adapt themselves to the representation and process of such reality.

For the necessities of representation, taking into account the structuring of biological knowledge [2], [5] is a progress that allows to consider useful common sense background knowledge in order to acquire, manage and process complex knowledge in a more elegant and efficient way.

The identification procedure that is described in this paper takes care of structured descriptions intelligently by reducing the number of eligible criteria for information gain calculation and manages coherent consultations through a guide to observation (web questionnaire).

Nevertheless, the problem we are faced with in Systematics is more difficult: good identifications depend on previously good classifications from experts, and also good descriptions from other biologists. Nature is so conceived that giving a name to organisms can also be difficult for experts (synonymies' problem), especially when there is a great intra-specific variation. This is the case in coral taxonomy where the number of named species in the world is uncertain [18]. Thus, managing complex knowledge in natural sciences means to cope with such evolving knowledge.

We have designed an Iterative Knowledge Base System to build knowledge bases in natural sciences that responds to these requirements. The main goal of IKBS is to produce quality descriptions which is a key factor for getting better results in identification process [11] and avoid future revisions.

2. Methodology

In the computer sciences, knowledge is a controversial term [Kodratoff, 1997]. We thus offer a working definition for our purpose in biology that consists of three kinds of knowledge: domain, instantiated and derived.

Domain knowledge (or background knowledge) relates to the definition of what is *observable*, i.e., build a descriptive model that corresponds to the modeling of data, or metadata [6]. *Instantiated knowledge* refers to the description of *observed* instances (case descriptions). *Derived knowledge* can be compared with produced hypotheses (cluster definitions, decision trees, rules, identification) discovered from domain and instantiated knowledge. Obviously, knowledge is also grounded in expert's mind and what is "extracted" is but a minimal part of his or her experience.

Knowledge Discovery methodology views knowledge as an output of a linear process of input data handling [8]. In biological domains, our emphasis is placed on a different interpretation of knowledge which consists of both input (domain and instantiated) and output (derived). This viewpoint is more relevant to Case-Based Reasoning methodology: i.e. the CBR cycle described in [1] with an extensive use of domain knowledge in the processing phase.

In practice, knowledge is extracted with IKBS by a cyclical process, divided into three parts:

1. Knowledge acquisition:
   - Acquire a descriptive model (domain knowledge or observable facts),
   - Acquire descriptions (observed facts or cases),
2. Knowledge processing:
   - Generate classification rules with decision tree induction,
   - Identify new observations(unknown specimens) with case-based reasoning,
3. Knowledge validation and refinement:
   - Verify the origin of misidentifications by analyzing differences of interpretation between the expert and the users of the knowledge base,
   - Iterate on the definition of the descriptive model (characters), update old cases.

For experts in biology, this approach is well suited to the natural process of their knowledge acquisition (conjecture and test) [16]:

1. Observe and familiarize oneself,
2. Represent observations, i.e. make descriptions,
3. Build hypotheses from descriptions (pre-classified), i.e. generate identification keys,
4. Test and experiment them with new observations, i.e. identify new specimens,
5. Refine their initial knowledge (new characters, cases and classifications).

The last point of the method is fundamental because the building of a knowledge base in natural sciences is very difficult. This was our experience in applications such as diagnosis of plant pathologies [12]. It is hard for experts to define the best representation of reality at once in a descriptive model. The challenge is to acquire the best character definitions and illustrations leading to interpretations of observations understood by anyone consulting the knowledge base.

3. Knowledge acquisition

This part of the methodology is very important to acquire a case base with good quality descriptions, i.e. that are well structured in different dimensions with all the required information, with characters, illustrations and comments that are easily comprehensible for other biologists.

3.1 The descriptive model

The descriptive model represents all the *observable* characteristics (objects, attributes and values) pertaining to individuals belonging to a particular domain. It is organized in a structured scheme, the name of the domain being at the root of a description tree. Each node of the tree is an object (a component of the individual) defined by a list of attributes with their respective possible values. Designing a descriptive model is essentially an expert task.

For helping them, we have set up logical rules for case description covering: decomposition, viewpoint, iteration, specialization, contextual conditions, etc. [11]. These rules were constructed from the analysis of expert's process of creating monographs of organisms or diseases.

To serve as an example, we present the descriptive model of one of the world's most widespread family of corals: *Pocilloporidæ* [7] (see Fig. 1). 51 objects and 120 attributes have been defined by the expert. With them, biologists are able to describe 4 genus and 14 species (see attribute called "taxon" in Fig. 1).

There are multiple benefits in such a representation. Viewpoints divide the descriptive model into homogeneous parts, thus giving a frame of reference for describing organisms at a particular level of observation (see objects identification, context, description, macro and micro structure in Fig. 1).

Sub-components introduce modularity into the descriptions making it possible to structure the domain from most general to most particular parts. This object representation of specimens is semantically better than the flat feature-value one: in the former, local descriptions of attributes depend on the existence of parent objects, although in the latter the defined characters are independent of one another. Some of the possibly missing objects are marked with a minus sign (e.g. columella).