FLEXIBLE AND FINE-GRAIN DECISION SUPPORT BASED ON DATA INTERPRETATION

T. Tamisier , O. Parisot, Y. Didry, J. Wax, F. Feltz

Centre de Rercherche Public - Gabriel Lippmann
41, rue du Brill, L-4422 Belvaux, Luxembourg
{wax,otjacque,tamisier,parisot,didry,feltz}@lippmann.lu

Decision support systems are nowadays used to disentangle intricate situations and perform sophisticated analysis [1]. However, there is often a gap between the characterization of a problem in terms of business knowledge and the computed answers restricted to a final decision result [2]. Through a long-time partnership with the administration of the Grand-Duchy of Luxembourg, the CRP - Gabriel Lippmann has been developing Cadral, an expert system that allows managing and formalizing extra information beside the mere decisional knowledge. This information can be useful to tune the knowledge model according to operational data, or to better exploit the decision result in the subsequent stages of a collaborative workflow.

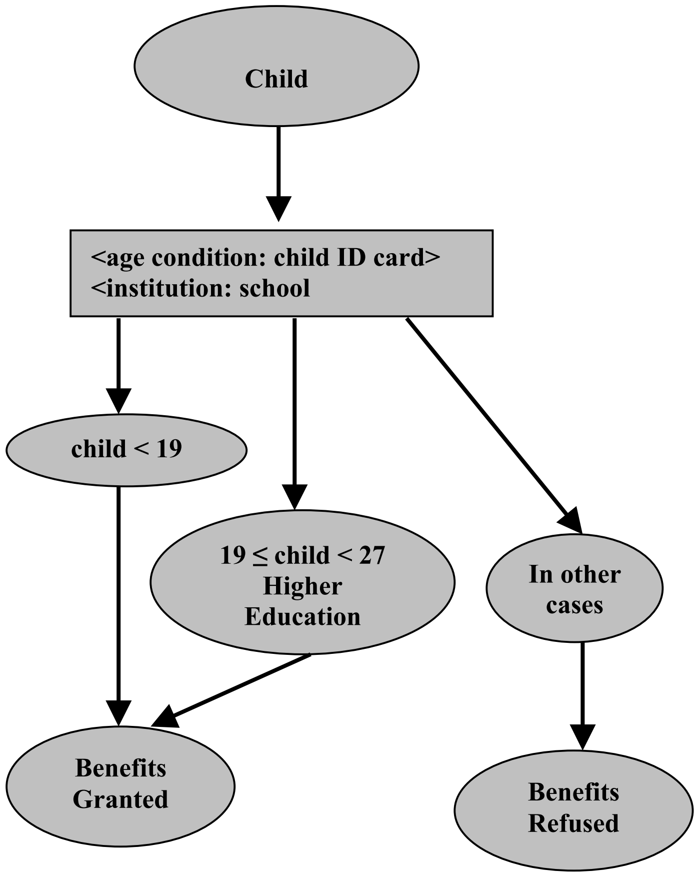
1. Cadral Architecture. The main application of Cadral is the processing the family benefits applications received by the Caisse Nationale des Prestations Familiales (National Family Benefits Fund) of the Grand-Duchy of Luxembourg. Handled administrative procedures are very complex due to the local open-economy where individual cases pertain to different national, supra-national and bilateral legal frameworks. Instead of the whole modeling of legal knowledge, Cadral is based on an explicit drawing of the mental procedures that governs the processing of the applications by interpreting operational data retrieved from the application files and centralized databases of the national administration [3].

Links to a legal database, used in connection with the reasoning trace of the system, allows exhibiting a legal justification of the resulting decisions. All the knowledge is therefore organized into 3 kinds of data: procedural data (processing rules), reference data (law), and application data (individual cases). The procedural knowledge manages metadata to retrieve and interpret the two other datasets. In particular, individual parameters relevant to a demand, including the presence of required certificates, are dispatched on several tables, and special care must be taken in order to recombine the information.

The choice of data tables as the primary format to store this heterogeneous information is justified to ease cooperative works between the Fund's collaborators for running and updating Cadral. In fact, the legal-based knowledge data is subject to frequent evolution, dismissal, or addition, that must be dealt with by non-computer specialized users. Cadral includes a knowledge editor that offers 2 different views to the procedural knowledge. In the analytical view, the knowledge is modeled on elementary if-then rules, which are directly processed by a resolution engine written in the Soar architecture [4]. The synthetic view offers a pictorial representation of all the knowledge, and in particular, shows the inter-dependence of the rules and their legal references. The editor allows checking the coherence of the procedural framework, which guarantees the equitability of individual decisions.

2. Running example. An example of a procedure can be: "If a child is going to school, and is less than 18, the beneficiary is entitled to receive the education allowance". Such a procedure is modeled in the shape of a multi-valued acyclic graph, with nodes representing a factual state (e.g. child going to school) used as a condition, and the edges denoting the necessary steps (e.g. showing a school certificate) to enter the state. Moreover, we ensure that a state is always unique in the graph (there are not 2 nodes with the same label), though it is fully possible to go to the same state by different ways. Let us consider the text of one law article. This text consists of several alineas (i.e. paragraphs), each of them associated with a different state. The modeling of the full law article is therefore a procedural graph, where all the states are distinct, and such that we can define an isomorphism, which associates every node in the graph with an alinea. When the procedures are translated into a collection of inference rules for an expert system, such an isomorphism is used concurrently with the trace of the inference engine in order to memorize the legal references made during the reasoning performed according to the procedures.

Figure 1 illustrates the graph-based modeling of the legal texts. The root node is associated with the allinea 1 of the (simplified) text, the two following node (in going down) with the alineas 2 and 3 respectively. Two additional concluding leaves for accepting or rejecting the applications show the final status.



Article: Simplified Process of Child Benefit
$1. A child benefit is granted to support the education of every child.
$2. The benefit is granted if the child is younger then 19.
$3. The benefit is granted, if the child is between 19 and 27 and follows universitary cursus.

Figure 1. Graph modeling of a simplified law article

3. Perspectives. With a view to better adapt the decision results to business requirements, we are now supplementing Cadral with a machine learning module. Three different modes for using the toolkit have been implemented: cluster building, relation setting, and decisional intelligence.

As a method to extract information from bunches of data, clustering aims at discovering similarities to group the data into different sets [5]. It has been successfully applied in decision-support to structure the raw data and optimize the application of solving strategies [6]. In Cadral, clustering is based on similarities drawn from key criteria selected by experts when modeling the rules. It is used for two improvements. First, for adapting the processing workflow to specific profiles requiring special care. In addition to processing application files, Cadral is also used for socio-economic prediction on projected demographic data or administrative and legal framework. The second use of clustering allows interpolating or correcting missing or unreliable data in predictive mode, as discussed in [7].

A variant of clustering, relation setting in Cadral adapts the processing workflow to input data based on pre-established criteria, calculated by the systems and hand-managed by the users. Relation setting is mainly used to refine the processing of specific input profiles (i.e. descriptive data relevant to benefit request files) requiring special care. For instance, Cadral manages a historisation module recording the previous computed results along with errors relevant to profiles' particulars. This historical information is then available to feed a relation base between profiles and possible processing errors.

Decisional intelligence is used by Cadral used to calculate the operational value of data not directly available (confidential, not recorded…). Manipulating this kind of data is quite different from manipulating missing data, because we never have values for them, so we can't use classical approximation or imputation, and we have to guess their values through mere "discovery" techniques. Among these techniques, neural networks have proven useful to derive an acceptation model from input and outputs datasets [8]. Hence, the machine learning toolkit allows building fictive data structures containing values of all available data, and uses neural networks to learn the particulars of data to discover.

References

[1] P. McCorduck, "Machines Who Think" (2nd ed.), A. K. Peters, 2008.

[2] T. Redman, "The impact of poor data quality on the typical enterprise", *Communications of the ACM*, 1998.

[3] T. Tamisier, and al, "A Collaborative Reasoning Maintenance System for a Reliable Application of Legislations", *Proc. of the 6th Int'l CDVE Conference*, LNCS, Springer, 2009.

[4] J. Laird, and al, "SOAR: an architecture for general intelligence", *Artificial Intelligence*, v. 33-1, 1987.

[5] V. Kumar, M. Steinbach, and P.-N. Tan, "Introduction to Data Mining", Chap. 8: Cluster Analysis, Addison-Wesley, 2006.

[6] L. Candillier, "Contextualisation, Visualisation et Evaluation en Apprentissage non Supervisé", *PhD thesis*, University of Lille-3, France, 2006.

[7] K. Lakshminarayan, S. Harp, R. Goldman, and T. Samad, "Imputation of missing data using machine learning techniques", http://www.aaai.org/Library/KDD/1996/kdd96-023.php

[8] Y. Wang, and H.-J. Xing, "Knowledge Discovery and Integration Based on A Novel Neural Network Ensemble Model". *Proc. of the International Conference on Semantics, Knowledge and Grid*, 2006.