Knowledge-Based Systems 67 (2014) 117–130

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Ontology-aware prediction from rules: A reconciliation-based approach

Fatiha Saïs [a], Rallou Thomopoulos [b,c,*]

[a] LRI, Université Paris-Sud, F-91405 Orsay cedex, France
[b] IATE Joint Research Unit, UMR1208, CIRAD-INRA-Supagro-Univ. Montpellier II, 2 place Viala, F-34060 Montpellier cedex 1, France
[c] INRIA GraphIK, LIRMM, 161 rue Ada, F-34392 Montpellier cedex 5, France

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form 17 April 2014
Accepted 28 May 2014
Available online 14 June 2014

Keywords:
Reasoning from knowledge
Information integration
Prediction
Case-based reasoning
Data reconciliation

ABSTRACT

Our work is related to the general problem of constructing predictions for decision support issues. It relies on knowledge expressed by numerous rules with homogeneous structure, extracted from various scientific publications in a specific domain. We propose a predictive approach that takes two stages: a reconciliation stage which identifies groups of rules expressing a common experimental tendency and a prediction stage which generates new rules, using both descriptions coming from experimental conditions and groups of reconciled rules obtained in stage one. The method has been tested with a case study related to food science and it has been compared to a classical approach based on decision trees. The results are promising in terms of accuracy, completeness and error rate.

1. Introduction

In the very last decades, extracting new knowledge from scientific publications has aroused great interest, in particular in experimental science domains, due to several converging circumstances and techniques: mass digitization of documents, web-enabled access to information, new experimental techniques allowing for high-throughput data acquisition, such as in genome sequencing for instance, but also new requirements for a higher and better-controlled production of goods. Indeed, the abundance of accessible scientific results both represents a real resource, and provides new needs for knowledge acquisition. This knowledge, once extracted from scientific publications, may be stored in a knowledge base. It can be exploited, among other uses, to answer user queries or to help for decision making issues.

However, one important problem of these knowledge bases is their incompleteness [1]. This incompleteness may be dealt: (i) by adapting the reasoning mechanisms for handling knowledge bases with omitted information [2]; (ii) by collecting new information from domain experts or from external sources like the World Wide Web [3]; or (iii) by using existing knowledge to predict unfilled information [4]. Our work falls in the third category. We propose a novel, case-based related approach for knowledge prediction that relies on reconciliation (which is a subfield of information integration).

Our application domain concerns food quality management in the cereal agrifood chain. Preliminary studies to this work were carried out on very different cases, outside the food science domain [5,6]. They have the following characteristics:

(1) The knowledge base is composed of a set of causality rules with homogeneous structure made up from a collection of scientific publications. They express syntheses of published experimental studies, obtained and validated through repeated experimentations. These rules are used for prediction. However, there is a huge number of possible experimental conditions. Consequently the knowledge base is incomplete by nature, since only a limited part of the possible experimental conditions have been explored in the literature and established as domain rules. Therefore, to make predictions concerning unexplored experimental conditions, a solution consists in using existing rules that concern close – although not identical – experimental conditions. In the more classic case where one starts from raw data, this approach is the principle of case-based reasoning.
(2) Although the rules concern distinct experimental conditions, they sometimes only differ by a small variation of one experimental parameter, which may be fundamental in the case of a highly discriminant parameter, but negligible for a parameter with low discriminance. Hence, rules which correspond

[*] Corresponding author at: INRA, IATE Joint Research Unit, 2 place Viala (bât. 31), F-34060 Montpellier cedex 1, France. Tel.: +33 4 99 61 22 17.

E-mail addresses: fatiha.sais@lri.fr (F. Saïs), rallou.thomopoulos@supagro.inra.fr (R. Thomopoulos).

http://dx.doi.org/10.1016/j.knosys.2014.05.023
0950-7051/© 2014 Elsevier B.V. All rights reserved.

to close experimental conditions and show similar results may be reconciled into groups of rules. Such groups have a semantics, since they express a common experimental tendency. They can also be exploited to reduce the search space in the prediction process. Performing this identification is a **reconciliation problem**. In addition to the experimental knowledge, general domain knowledge is available, and it has been modeled in an ontology. The ontology includes a vocabulary organized by subsumption, disjunction and synonymy relations. Moreover, it provides less common information concerning the status of concepts, such as functional dependencies and discriminance of concepts for prediction. Several existing methods aim at evaluating the similarity of data descriptions for various purposes (e.g. for prediction in case-based reasoning, for grouping into classes in classification, for detecting whether different data refer to the same real world entity in data reconciliation). Methods from data reconciliation [7,5] are the most advanced ones: they take into account the logical semantics of an ontology, with particular attention to the functional dependencies.

The objective of this paper is to propose an approach to generate prediction rules relying on case-based and reconciliation methods, using an ontology. The approach we propose performs two stages that exploit the ontology:

- rule reconciliation into groups that express common experimental tendencies. From a mosaic of isolated pieces of knowledge, we identify the main experimental zones, which is also the experts' way of proceeding while analyzing an experimental domain of knowledge;
- computation of a prediction rule, starting from a new description of experimental conditions and from the "closest" group of reconciled rules.

Our method has been tested within a food science application concerning food quality management in the cereal agri-food chain and it has been compared to a classic predictive technique, using decision trees.

Not every predictive method may be used in the considered context. Experimental conditions have **missing values** (not all the parameters are described in each rule), use **both quantitative and qualitative** parameters (numerical and symbolic values), and are **scarce** since we are not in a high speed data application but in a scarce-knowledge context with only a few hundreds of rules available. Few methods are able to deal with all these three issues and they basically are case-based approaches or decision trees methods. This is the reason why we decided to compare our work against these approaches.

The paper is organized as follows: Section 2 describes the formalism used for ontology and domain rules representation. Section 3 gives an overview of related work in case-based reasoning, data reconciliation and decision tree prediction. Section 4 is dedicated to the proposed rule reconciliation method. Section 5 presents the proposed prediction method. Section 7 describes the context and related work in the application domain and proposes a comparative evaluation of the developed approach. Finally, Section 8 concludes the paper by giving some future work perspectives.

2. Preliminaries

In this section, we briefly recall essential elements regarding ontology and domain rule definition.

2.1. The domain ontology

The ontology $\mathcal{O}$ is defined as a tuple $\mathcal{O} = \{\mathcal{C}, \mathcal{R}\}$ where $\mathcal{C}$ is a set of concepts and $\mathcal{R}$ is a set of relations.

2.1.1. Ontology concepts

Each concept c is associated with a definition domain by the *def* function. This definition domain can be:

- *numeric*, i.e. $def(c)$ is a closed interval $[min_c, max_c]$;
- *'flat' (non hierarchized) symbolic*, i.e. $def(c)$ is an unordered set of constants, such as a set of bibliographic references;
- *hierarchized symbolic*, i.e. $def(c)$ is a set of partially ordered constants, themselves are concepts belonging to $\mathcal{C}$.

In the sequel, we will refer to elements of concept domain definition by *values*.

2.1.2. Ontology relations

The set of relations $\mathcal{R}$ is composed of:

- the *subsumption* or 'kind of' relation denoted by $\preceq$, which defines a partial order over $\mathcal{C}$. Given $c \in \mathcal{C}$, we denote as $\mathcal{C}_c$ the set of sub-concepts of c, such that: $\mathcal{C}_c = \{c' \in \mathcal{C} | c' \preceq c\}$. When c is defined by hierarchized symbolic definition domain, we have $def(c) = \mathcal{C}_c$.
- the *equivalence* relation, denoted by $\equiv$, expressing a synonymy between concepts of the ontology.
- the *disjunction* relation between concepts, denoted by $\perp$. Given two concepts c and $c' \in \mathcal{C}, c \perp c' \Rightarrow (def(c) \cap def(c')) = \emptyset$. We note that the disjunction relation respects the subsumption relation. This means that if two general concepts c and c' are declared as disjoint then all the concepts that are more specific than c and c', respectively, are pairwise disjoint.

Fig. 1 gives a small part of the set of concepts $\mathcal{C}$, partially ordered by the subsomption relation (pictured by '→'). Examples of disjunctions are given apart for readability reasons. Note that the considered ontologies are not restricted to trees, they are general graphs. This is an important feature of our work with respect to previous approaches, such as [8], where only trees are considered.

2.1.3. Least common subsumer

Given two concepts c_1 and c_2, we denote as $lcs(c_1, c_2)$ their least common subsumer, that is $lcs(c_1, c_2) = \{c \in \mathcal{C} | c_i \preceq c$, and $((\exists c'$ s.t. $c_i \preceq c') \Rightarrow (c \preceq c')), i \in \{1, 2\}\}$.

For example, in the ontology of Fig. 1, the *lcs* of the concepts *LiposolubleVitamin* and *VitaminB* is the concept *Vitamin*. As commonly done, we consider a *Universal* concept subsuming all the other concepts of the ontology, to ensure that such a *lcs* always exists.

2.1.4. Relationship between ontology concepts and experimental variables

We consider a set of experimental descriptions containing K variables. Each variable $X_k, k = 1, \ldots, K$, is associated with a concept $c \in \mathcal{C}$ of the ontology $\mathcal{O}$. Each variable can be instantiated by a value that belongs to the definition domain of concept c.

2.1.5. Variable discriminance

For each variable X_k, a discriminance score, denoted by λ_k, is declared. It is a real value in the interval $[0; 1]$. It is obtained through an iterative approach performed with domain experts, as briefly explained in Section 4.3 and in more detail in [9].