Knowledge-Based Systems 92 (2016) 183–199

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A novel method for rule extraction in a knowledge-based innovation tutoring system

Harold Paredes-Frigolett[a,*], Luiz Flávio Autran Monteiro Gomes[b]

[a] Faculty of Economics and Business, Management Department, Diego Portales University, Av. Santa Clara 797, Santiago, Chile
[b] IBMEC School of Business and Economics, Av. Presidente Wilson 118, 20030-020 Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 17 February 2015
Revised 8 October 2015
Accepted 9 October 2015
Available online 10 November 2015

Keywords:
Knowledge-based systems
Knowledge acquisition
Knowledge representation
Innovation management
Multicriteria decision analysis

ABSTRACT

Symbolic, subsymbolic, and hybrid approaches to rule extraction have so far relied on subsets of first-order logic to cope with the expressiveness trade-off of knowledge representation, on black-box approaches based on artificial neural networks, or on frequent association rule mining in the knowledge discovery and data mining fields. In this article, we present an entirely new method for rule extraction in knowledge-based systems that consists in retrieving an initial set of rules extracted from a knowledge base using conventional logical approaches and then ranking this initial set of rules applying a psychologically motivated multicriteria decision analysis method. We show how this method can be used to implement a knowledge-based management system, demonstrate that this method outperforms the most efficient algorithms for rule extraction proposed to date in the knowledge representation and knowledge discovery fields, and describe its implementation in a knowledge-based innovation tutoring system.

1. Introduction

In the field of Artificial Intelligence there has been a long debate concerning the need for a logical foundation of the knowledge representation language to be used in order to implement commercial-grade, knowledge-based systems. Initial approaches to the problem of knowledge representation postulated a so-called operational semantics, that is, an *ad hoc* semantics not based on formal logics but on the inner workings of the knowledge representation system *per se*, hence the name operational semantics [59,62,64]. Operational semantics was often difficult to formalize and, more importantly, was not well-suited for reasoning in a logically verifiable sense.

During the 1980s and 1990s a discussion ensued in the AI community regarding the need for a logical foundation in a knowledge representation system. As a result, several knowledge representation languages were put forward based on first-order predicate logic [8,9,12]. By the late 1980s, the apparent need for a logical foundation in a knowledge representation system was undisputed in the knowledge representation community [35]. This led to the second – and still unsettled – debate on how much expressiveness was required, that is, how powerful the logical foundation of the knowledge-based system and its associated knowledge representation language should be [80].

With the development of knowledge representation systems in the tradition of KL-ONE [14,18], which were based on the formalism of Semantic Networks [59,84,85] and postulated a subset of first-order predicate logic as the underlying logical foundation of a knowledge representation system, the need for restraining the expressiveness of the underlying semantic representation language became the mainstream approach in the knowledge representation community. This shift toward constraining the expressiveness of the logic was seen as a necessary requirement in order for the underlying inferential algorithms to show good tractability behavior [55]. This trend toward reducing expressiveness in favor of tractability was embraced by most knowledge representation systems based on restricted subsets of first-order predicate logic [3,4,67,83]. As mentioned, this debate has not yet been settled, as there is a difficult trade-off between expressiveness in the knowledge base, on the one hand, and tractability of the underlying inferential algorithms, on the other [88]. This problem of expressiveness is a core problem in the field of knowledge representation, particularly when it comes to the development of knowledge-based systems for real-life applications such as managing processes of innovation and entrepreneurship. Expressiveness in this context is defined as the capability of the knowledge representation language to represent knowledge of the application domain in a way that is computationally tractable and semantically well-defined. Unfortunately, most application domains in the area of management are very knowledge intensive, thus requiring the deployment of large knowledge bases and the use of more expressive knowledge representation languages. As a result, very few knowledge-based

* Corresponding author. Tel.: +56 2 22130127.
E-mail addresses: harold.paredes@udp.cl, hparedesfrigolet@hotmail.com (H. Paredes-Frigolett), autran@ibmecrj.br (L.F.A.M. Gomes).

http://dx.doi.org/10.1016/j.knosys.2015.10.027
0950-7051/

approaches to evidence-based management have been attempted, often for very restricted application domains, with very few exceptions of large infrastructure projects.

The motivation for the work reported in this article has been the development of efficient algorithms for rule extraction from large knowledge bases to be used in concert with an evidence-based innovation tutoring system designed to bridge the knowledge gaps of emerging regions of innovation and entrepreneurship. The adequacy of such an approach to solving this problem depends upon satisfying the following three conditions [80]:

1. *Knowledge representation*: Defining and implementing a common knowledge representation language and system for representing knowledge and reasoning in a given application domain (innovation and entrepreneurship in knowledge-intensive industries in our case).
2. *Knowledge acquisition*: Acquiring knowledge from a variety of structured and semi-structured sources to generate large knowledge bases containing the knowledge required for a given application domain (the knowledge that is diffused in emerging and complex innovation networks in our case).
3. *Knowledge extraction*: Providing tools for automatic extraction of knowledge rules from knowledge bases so that relevant answers and/or recommendations can be generated and presented to users (rules about processes of innovation and entrepreneurship in our case).

The contribution of this article is a new and highly efficient method to solve the third problem above, namely, the automatic extraction of rules from a knowledge base in knowledge-based management systems. The resulting rule extraction mechanism has been designed to be used in concert with a knowledge-based tutoring system and builds upon a series of functions that measure the relevance of rules in a knowledge base. These relevance functions draw upon a comprehensive body of work on three main fronts: the logical foundations of Artificial Intelligence [8,35], knowledge-based systems and their associated knowledge representation systems[3,4,52–55,80,84–89], and knowledge discovery and data mining [1,50,61,97].

The article is organized as follows. In Section 2, we present previous work in the area of knowledge extraction. In Section 3, we present how the problem of knowledge extraction from large knowledge bases can be modeled as a multicriteria decision analysis problem. In Section 4, we present the multicriteria decision analysis method used to implement the rule extraction mechanism based on the problem definition presented in Section 3. In Section 5, we provide a case that shows how the method for rule extraction introduced in Section 4 is used in a first implementation of a knowledge-based innovation tutoring system so as to comply with the systemic criteria of scalability and performance required for this system. We also present the complexity of our method and show that it outperforms the most efficient algorithms for rule extraction in the fields of knowledge representation and knowledge discovery. In Section 6, we present the discussion of our results. In Section 7, we present our conclusions.

2. Previous work

The problem of knowledge extraction has traditionally been dealt with in the field of knowledge-based systems by choosing less expressive knowledge representation languages. The mainstream approach to building large knowledge representation systems has consisted in reducing the expressiveness of the underlying knowledge representation language so as to be able to implement efficient algorithms for rule extraction and reasoning [55]. Unfortunately, even for highly restricted subsets of first-order predicate logic such as description logics [3,4], terminological logics [14], and typed-feature logics [83], the reasoning algorithms become quite intractable as the number of rules in the knowledge base grow [68].

Conventional approaches to solving this problem have traditionally relied on standard algorithms implemented in knowledge representation systems such as subsumption in description logics [3,4,67] and unification in typed-feature logics [83]. More expressive knowledge representation systems that embrace the full representational power of first-order predicate logic rely on algorithms for forward (input-driven) or backward (goal-driven) inference [80,81]. The latter approaches disregard the expressiveness trade-off of knowledge representation, that is, the fact that more expressive knowledge representations will lead to more inefficient – and even undecidable – algorithms for reasoning. Despite the important initial advances in the area of knowledge extraction, much work still needs to be done in order to build and extract rules from large knowledge bases for commercial-grade applications using symbolic approaches. Large knowledge infrastructure projects such as the CycL project [53,54] demonstrated that this endeavor is indeed possible for commercial-grade applications in specific domains. More recent advances in extracting knowledge from texts in the symbolic tradition such as the work done in connection with the KNEXT project at the University of Rochester promise to provide us with tools for automatic knowledge extraction from semi-structured and noisy text available on the Web today [24–26,42,79,80,82].

In more recent years, the rise of non-symbolic approaches such as artificial neural networks [32,33,78] contributed to the emergence of so-called hybrid approaches to knowledge extraction [44,56,63]. The emergence of these hybrid approaches was fueled by the initial success of connectionist approaches [57] despite the fact that these black-box approaches fail to provide the level of explanation that only white-box symbolic approaches can deliver for complex application domains [15,27,28]. Due to the shortcomings of the rule extraction approaches based on neural networks [47], it became clear that traditional white-box, symbolic approaches based on logics needed to be used in concert with these connectionist, subsymbolic approaches based on neural networks in order to solve the so-called theory and knowledge bottleneck of knowledge-based systems [80]. As a result, further integration of expert systems based on artificial neural networks with other formalisms such as fuzzy logic [99] and rough sets [72] has contributed to the emergence of a wide variety of hybrid expert systems in different application domains [2,19,20,23,45,60].

Closely related with the rule extraction method proposed in this article are a series of methods in the area of association rule mining [1,50,61,97]. These approaches to rule extraction intend to extract a set of nonredundant frequent rules over a given threshold of "interestingness" from a database of transactions consisting of uniquely identified subsets of items taken from a predefined set of items [1]. Frequent association rule mining is an important area of data mining that has suffered from the fact that the association rules mined produced extremely large sets of mostly uninteresting rules [97]. Most of the work in these areas has focused on reducing the search space in order to come up with efficient rule extraction algorithms for complex commercial-grade applications based on the rule extraction algorithms proposed for the less complex basket market analysis applications originally investigated in connection with frequent association rule mining [1,50,61,97]. Our approach also departs from probabilistic approaches in the tradition of belief rule-based systems [21,49,90,103] in that our application domain does not call for heavy-duty inferential processes in the knowledge base. In our application, probabilistic conditionals stored in the knowledge base are used to encode generic knowledge about the world and the application domain of innovation management. They are used to infer new rules over an interestingness threshold and are not intended to implement the kind of predictive models typically associated with belief rule-based systems. Contrary to belief rule-based systems, our application domain requires a more straightforward way of calculating