Information Sciences 194 (2012) 107–119

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Knowledge discovery through ontology matching: An approach based on an Artificial Neural Network model

M. Rubiolo *, M.L. Caliusco, G. Stegmayer, M. Coronel, M. Gareli Fabrizi

National Council of Scientific and Technical Research (CONICET), Information System Engineering Research and Development Center (CIDISI), Technological National University (UTN-FRSF), Argentina

ARTICLE INFO

Article history:
Available online 24 August 2011

Keywords:
Semantic Web
Artificial Neural Network
WordNet
Knowledge-source discovery

ABSTRACT

With the emergence of the Semantic Web several domain ontologies were developed, which varied not only in their structure but also in the natural language used to define them. The lack of an integrated view of all web nodes and the existence of heterogeneous domain ontologies drive new challenges in the discovery of knowledge resources which are relevant to a user's request. New approaches have recently appeared for developing web intelligence and helping users avoid irrelevant results on the web. However, there remains some work to be done. This work makes a contribution by presenting an ANN-based ontology matching model for knowledge source discovery on the Semantic Web. Experimental results obtained on a real case study have shown that this model provides satisfactory responses.

1. Introduction

Currently, the web consists of a large number of distributed and connected texts. These texts, however, are neither understood by software applications nor really used by them. The key for solving this problem is to improve both the ability of software applications to communicate directly with each other, and the representation of information in a way far more usable by them. An important framework for creating such capabilities can be provided by the next generation of the Web architecture: the *Semantic Web* [30]. The Semantic Web is intended to be a paradigm shift just as powerful as the original Web. It will bring meaning to Web page contents, in which software agents roaming from page to page can execute automated tasks by using metadata or semantic annotations, ontologies and logic [1].

Ontologies have become a de facto standard for the semantic description of data, resources and services in large distributed systems such as the Web [2]. It has been predicted that within the next decade, the Web will be composed of small highly contextualized ontologies, developed with different languages and different granularity levels [19,20]. In this context, the simplest document will consist of concrete facts, classes, properties definitions and metadata [23]. The websites will have not only a domain ontology to describe the knowledge they can provide, but also an adequate structure to receive mobile software agents that will travel through the net (for example, looking for knowledge required by an end-user) [30].

Although the capabilities and scope of the Semantic Web are impressive today, its continuous evolution presents many problems to be faced. For instance, whenever a node on the Web needs to initiate a dynamic collaborative relationship with another, it certainly finds it difficult to know which node to contact or where to look for the required knowledge. Therefore, it can be seen that the knowledge resource discovery in such an open distributed system becomes a major challenge. This is due to the lack of an integrated view of all the available knowledge resources.

Besides, the existence of multiple domain-specific heterogeneous ontologies and their distributed development introduces another problem: on the Semantic Web, many independently developed ontologies describing the same or very

* Corresponding author.
E-mail address: mrubiolo@santafe-conicet.gov.ar (M. Rubiolo).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2011.08.008

similar fields of knowledge, co-exist. These ontologies are either non-identical or present minor differences, such as different naming conventions or different structures in the way they represent knowledge. Any application that involves multiple ontologies must establish semantic mappings among them to ensure interoperability. Examples of such applications arise in many domains, including e-commerce, e-learning, information extraction, bioinformatics, web services, tourism, among others [15]. For that reason, ontology-matching is necessary to solve the problem.

Ontology-matching techniques, essentially, identify semantic affinity between concepts belonging to different ontologies. Among recent proposals, machine learning methods can process the matching problem through the presentation of many correct (positive) and incorrect (negative) examples. Such algorithms require sample data from which to learn. Matchers using machine learning usually operate in two phases: (i) the learning or training phase and (ii) the classification or matching phase. During the first phase, training data for the learning process of a matcher is created, for example by manually matching two ontologies. During the second phase, the trained matcher is used for matching new ontologies. Learning can be processed on-line – the system can continuously learn – or off-line, so its speed is not as relevant as its accuracy. Usually, this process is carried out by dividing the available data set. For instance, considering a set of positive and negative examples of alignments, the examples would be divided into a training set (typically 80% of data) and a validation or test set (typically 20% of data), which would be used for evaluating the performance of the learning algorithm [17].

This paper presents a approach for improving web knowledge resource discovery on the Semantic Web based on recently developed intelligent techniques. This approach combines agent-based technologies with a machine-learning classifier (in particular an Artificial Neural Network (ANN) model) used for defining the matching operation between ontologies, and makes a proposal for codifying ontology knowledge as inputs to the neural model. The method proposed in this paper takes into account both schema-level and instance-level information from ontologies and semantic annotations, and proposes a machine learning solution to the ontology-matching problem, which has been proved in a real world case study, obtaining high matching rates.

The paper is organized as follows. Section 2 introduces some issues related to ontology-matching and ANN concepts. Section 3 presents related work. In Section 4, a motivating scenario for knowledge discovery is introduced, and an intelligent web-agent is presented. Section 5 explains the ANN-based ontology-matching model used by this agent. Section 6 presents the model evaluation by considering a case study. Section 7 compares the ANN-based model with the H-Match ontology matching algorithm. Finally, conclusions and future work can be found in Section 8.

2. Preliminaries

The purpose of this section is to provide definitions of the terminology used in this work.

2.1. The role of ontologies and software agents

Ontologies provide a number of useful features for intelligent systems, as well as for knowledge representation in general, and for the knowledge engineering process [19]. In the literature, there are different definitions about what an ontology is and its different classifications [21,20]. In this paper, a *domain ontology* is considered a representational artifact of the semantics of a given domain of discourse. A *domain* is a slice of reality that forms the subject-matter of a single science or technology [31]. The representational units of an ontology are the following: *terms*, *relations*, *properties*, *axioms* and *instances*. Since domain ontologies provide a shared conceptualization of a certain domain, they can be used for describing resource semantics by adding metadata to them.

In order to define the semantics for a digital content, it is necessary to formalize the ontologies by using specific languages such as Resource Description Framework (RDF) and Web Ontology Language (OWL) [32]. While RDF is a general-purpose language for representing information about resources in the Web, OWL is a semantic markup language for publishing and sharing ontologies. Although RDF was originally meant to represent the metadata of web resources, it can also be used to represent information about objects that can be identified on the Web. The basic construction in RDF is an (*Object*, *Property*, *Value*) triplet: a subject *S* has a property *P* with value *V*. An RDF-triplet corresponds to the relation that could be written as (*S*, *P*, *V*), for example (http://www.books.org/ISBN0012515866, has Price, 62); and (*Professor*, *teachesSubject*, *Artificial Intelligence*) in the case of a University website annotated with an ontology.

Besides ontologies, software agents will play a fundamental role in building the Semantic Web of the future [23]. When data is marked up using ontologies, software agents can understand the semantics better and, therefore, locate and integrate data more intelligently, for a wide variety of tasks. A new emergent category of agents, named intelligent web (or personal) agents, would find more possible ways to meet the needs of a user and offer different choices to accomplish their goals. A personal software agent on the Semantic Web must be capable of receiving tasks and preferences from a user, seeking information from heterogeneous web sources [5].

2.2. Ontology matching

Ontology-matching aims at finding correspondences between semantically related elements of different ontologies. The correspondences may stand for equivalence as well as other relations, such as consequence, subsumption, or disjointness,