Data & Knowledge Engineering 69 (2010) 573–597

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

A methodology to learn ontological attributes from the Web

David Sánchez *

Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA), Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Avda. Països Catalans, 26. 43007 Tarragona, Spain

ARTICLE INFO

Article history:
Received 21 January 2009
Received in revised form 25 January 2010
Accepted 27 January 2010
Available online 4 February 2010

Keywords:
Ontology learning
Meronyms
Attributes
Features
Web mining
Knowledge acquisition

ABSTRACT

Class descriptors such as attributes, features or meronyms are rarely considered when developing ontologies. Even WordNet only includes a reduced amount of *part-of* relationships. However, these data are crucial for *defining* concepts such as those considered in classical knowledge representation models. Some attempts have been made to extract those relations from text using general meronymy detection patterns; however, there has been very little work on learning expressive class attributes (including associated domain, range or data values) at an ontological level. In this paper we take this background into consideration when proposing and implementing an automatic, non-supervised and domain-independent methodology to extend ontological classes in terms of learning concept attributes, data-types, value ranges and measurement units. In order to present a general solution and minimize the data sparseness of pattern-based approaches, we use the Web as a massive learning corpus to retrieve data and to infer information distribution using highly contextualized queries aimed at improving the quality of the result. This corpus is also automatically updated in an adaptive manner according to the knowledge already acquired and the learning throughput. Results have been manually checked by means of an expert-based concept-per-concept evaluation for several well distinguished domains showing reliable results and a reasonable learning performance.

1. Introduction

Ontologies have emerged in recent years as a fundamental tool for formalizing and representing knowledge. They offer a formal and explicit specification of a shared conceptualization. With the massive growth of the information society and the success of the Web 2.0, the need for this kind of knowledge formalization model has become imperative. In fact, ontologies are a fundamental element for the success of the Semantic Web [68]. However, the ontological construction of such structures is typically carried out by knowledge engineers and domain experts, resulting in long and tedious development stages. Given the massive scope of the Semantic Web, the manual approach is not scalable enough. Because of this knowledge representation bottleneck, researchers have put their efforts into aiding the ontology construction process [56].

Ontologies are composed of at least three elements: *classes* (concepts of the domain), *relations* (different types of binary associations between concepts or data-values) and *instances* (real world individuals). Formally [56], an ontology is presented as an object model comprising a set of concepts or C classes which are *taxonomically* related by the transitive *is-a* relation $H \epsilon\ C\ x\ C$ (e.g. *dog is a mammal*) and *non-taxonomically* related by named object relations $R^* \epsilon\ C\ x\ C\ x\ String$ (e.g. *cigarettes cause lung cancer*).

* Tel.: +34 977 556563; fax: +34 977 559710.
E-mail address: david.sanchez@urv.net

0169-023X/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2010.01.006

From the point of view of automatic ontology learning, many approaches have been developed to acquire mainly domain concepts and organize them into taxonomies (as detailed in [56]). However, the identification of non-taxonomic relations has received very little attention [10,13].

Within the non-taxonomical field, we can identify special binary relations among concepts (*object relations*), which express *part-of* and associations between objects and data-values (*data-properties*), which are typically referred as *attributes* [1], *features* [8] or *parts* [39]. The former can be expressed as $P^* \in C \times C \times Part\text{-}of$ where C are classes of the ontology and *Part-of* expresses a *meronym* relationship between concepts (e.g. $p = Digital\_Camera \times CCD\_Sensor \times part\_of$). The latter can be expressed as $D^* \in C \times T \times String$, where C is a class of the ontology, T are data-types and *String* is a textual representation of the relation (e.g. $d = Digital\_Camera \times Float \times Price$). These relations (whose acquisition represents the central point of this paper) play a very important role in adding semantic content to ontological classes due to their descriptive nature. In fact, in classical theories of knowledge representation, concepts are defined in terms of their attributes (e.g. *color, shape, size*, etc.) [64]. Almuhared and Poesio [1] demonstrated that identifying a concept by its attributes leads to a better lexical description. Therefore, this data represents a valuable aid in knowledge driven tasks such as question answering [67], information retrieval [6] or word-sense disambiguation [2]. However, an investigation of the structure of existing ontologies via the *Swoogle* ontology search engine [37] has shown that available ontologies rarely model these kinds of relationships.

From the point of view of ontology learning, the following tasks should be performed to acquire expressive part-of object relations and data-properties: (i) discovery and labelling of relevant properties for a domain and, for data-properties, (ii) identification of the appropriate data-type and specification of possible value restrictions. Due to the generality and unbounded nature of literal data-values and the inherent ambiguity of human natural language, these are challenging tasks [39]. In fact, as will be shown in the related work section, even though some approaches in the field of Information Extraction have been developed to extract features, very little work has been done on acquiring expressive class attributes and restrictions in the field of ontology learning.

In this paper we present a new methodology for acquiring class attributes at an ontological level. In addition to object relations, one of the paper's contributions is to address the discovery of data-properties and their associated data-types and value ranges. Moreover, unlike many previous approaches, the method has been designed in an automatic and domain-independent way, exploiting several well-established analytical techniques. In order to minimize data sparseness which characterizes approaches based on the analysis of concrete and/or domain-dependant repositories, the Web is exploited as a social scale learning source. Due to the unsupervised nature of the employed analytical techniques, the method relies on the Web information distribution in order to assess the reliability of the extracted knowledge. Specially designed statistical scores based on collocation measures and highly contextualized assessments have been also designed in order to improve the accuracy of the results. The method has been manually evaluated for several well distinguished domains, showing its feasibility in obtaining relevant and reliable results in a scalable manner.

The rest of the paper is organized as follows. Section 2 presents an overview of previous approaches to learning concept attributes/features/meronyms from textual documents. Section 3 introduces the basis of our methodology, including a description of the main techniques employed to acquire and filter attribute candidates and a study of the Web as learning source. Section 4 describes in detail the proposed methodology, which is divided into a three-staged procedure and covers the acquisition of attributes, data-types and value ranges. Section 5 discusses some relevant aspects regarding the analysis of web resources and presents an adaptive algorithm for incremental corpus analysis. Section 6 describes the evaluation procedure and presents and discusses the results of several tests. Section 7 analyses the computational complexity of the proposed algorithms and shows the throughput and the practical feasibility of the methodology. The final section presents the conclusions and proposes some lines of future work.

2. Related work

The notion of concept *attribute* is not completely clear and the term has been used in widely different ways in knowledge representation literature [1]. Guarino [48] classified attributes into *relational* (e.g. color, position) and *non-relational* ones (like object parts). In the qualia structure of the generative lexicon [30] four types of roles are identified: *Constitutive Role* (parts), *Formal Role* (qualities), *Agentive Role* (relational) and *Telic Role* (purpose).

An analysis of previous research in the NLP literature on information extraction also shows different ways of referring to object–data–value relationships and how these fall into one or several of the previously stated categories. Typically, these relationships are considered as concept *attributes* [1] or *features* [8]. This means defining a certain data-type or measure range. Other authors refer to special object–object relations and talk about *meronyms* [45] or *part-of* [39]. In both these cases, the analysis only covers the discovery of the relation.

In the present study, we generally refer to *attributes* as:

Definition 1 (*attribute*). *Attributes* are object–object part-of relationships and object–data–value properties which can help to semantically define and describe an ontological concept.

Therefore, this definition ranges from pure part-of relationships (e.g. the *optical lens* of a *digital camera*) which can be represented as special object–object relationships, to specific features (e.g. *ISO* or *resolution* of a *digital camera*) and properties (e.g. *size* or *weight*) which can be qualified or numerically quantified. In the last case, attribute data-types, values and