Chapter 25
A Web-Based Method for Ontology Population

Hilário Oliveira
Federal University of Pernambuco, Brazil

Rinaldo Lima
Federal University of Pernambuco, Brazil

João Gomes
Federal University of Pernambuco, Brazil

Fred Freitas
Federal University of Pernambuco, Brazil

Rafael Dueire Lins
Federal University of Pernambuco, Brazil

Steven J. Simske
Hewlett-Packard Labs, USA

Marcelo Riss
Hewlett-Parckard do Brasil, Brazil

ABSTRACT

The Semantic Web, proposed by Berners-Lee, aims to make explicit the meaning of the data available on the Internet, making it possible for Web data to be processed both by people and intelligent agents. The Semantic Web requires Web data to be semantically classified and annotated with some structured representation of knowledge, such as ontologies. This chapter proposes an unsupervised, domain-independent method for extracting instances of ontological classes from unstructured data sources available on the World Wide Web. Starting with an initial set of linguistic patterns, a confidence-weighted score measure is presented integrating distinct measures and heuristics to rank candidate instances extracted from the Web. The results of several experiments are discussed achieving very encouraging results, which demonstrate the feasibility of the proposed method for automatic ontology population.

INTRODUCTION

In the last decades the amount of information generated in digital form and published on the Web has been growing daily at a fast rate. Nowadays, the Web can be considered as the largest information repository in the world, becoming a "library" of unprecedented size in human history, encompassing all domains of knowledge. Most of such knowledge is represented in textual format, written in natural language, and interpretable only by humans. Despite the increasing volume of data available on the Web, human capacity for processing and absorption of information remains

DOI: 10.4018/978-1-4666-8751-6.ch025

constant and limited (Ben-Dov & Feldman, 2005). In such context, it is of paramount importance to have computational systems that would automatically process and classify the huge volume of data available on the Internet.

Storing information in plain text format does not allow document accessibility, since the semantic aspects of its content are not explicitly expressed. The lack of some kind of structure hinders the exploration and interpretation of semantic information by computational agents (Fensel *et al.,* 2002). According to Feldman and Sanger (2007), the Web is currently at a syntactic level, i.e., its contents can be read by machines, just considering keywords and combinations of these, and not on a semantic level, in which computer systems can interpret unambiguously the information available. This characteristic constitutes an important limitation of the current Web. Thus, the task of automatically finding relevant information to specific needs, especially those that require some level of semantic interpretation, becomes arduous, costly, time consuming, and, in many cases impractical.

To address those limitations, the Semantic Web (Berners-Lee & Hendler & Lassila, 2001) was proposed as a global initiative, defined what would be the evolution of the current Web scenario. The main goal of the Semantic Web is to make explicit the meaning of the content of the data on the Web. Thus, it is possible that web data be processed by both people and computational agents which would have access to the semantic aspects of the data. The Semantic Web is based on a layered architecture in which each layer adds a higher level of expressiveness and inference (Koivunen & Miller, 2001) to the others. One of the fundamental layers in the development of the Semantic Web is composed by ontologies, which are responsible for providing the necessary expressiveness to the representation of relevant knowledge about a domain (Freitas, 2003). Thus, the first step to make the Semantic Web goals achievable is the definition of appropriate semantic structures for representing any possible domain of knowledge, which implies in the development of domain or task-specific ontologies. Once the ontology for a specific domain is available, the next step is to semantically annotate related web resources. Thus, computers must have access to ontologies that enable both the representation and sharing of knowledge of different domains, and a process for mapping the chosen ontologies to the web content.

On the other hand, although domain or task-based ontologies are recognized as essential resources for the Semantic Web, the development of such ontologies relies on domain experts or knowledge engineers that typically adopt a manual construction process. Such manual construction process is very time-consuming and error-prone (Cimiano, 2006). An automated or semi-automated mechanism to convert the information contained in existing web pages into ontologies is highly desirable. Ontology-based Information Extraction (OBIE) (Wimalasuriya & Dou, 2010), a subfield of Information Extraction (IE), is a promising candidate for such a mechanism. An OBIE system can process unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information, and present the output using ontologies.

Trying to overcome such a problem, this chapter proposes an unsupervised, automatic and domain-independent method able to assist the process of Ontology Population. The proposed methodology is able of extracting instances of ontological classes from unstructured sources of information written in natural language available on the Web. The method is driving by an input ontology that defines concepts which must be populated, and an initial set of linguistic patterns (seed patterns) used to extract and classify candidate instances. It is based on a Confidence-weighted Score function (*ConfScore*), which integrates different measures and heuristics to rank candidate instances.