Chapter 12
Evolutionary Conceptual Clustering Based on Induced Pseudo-Metrics

Nicola Fanizzi
Università degli studi di Bari, Italy

Claudia d’Amato
Università degli studi di Bari, Italy

Floriana Esposito
Università degli studi di Bari, Italy

ABSTRACT

We present a method based on clustering techniques to detect possible/probable novel concepts or concept drift in a Description Logics knowledge base. The method exploits a semi-distance measure defined for individuals, that is based on a finite number of dimensions corresponding to a committee of discriminating features (concept descriptions). A maximally discriminating group of features is obtained with a randomized optimization method. In the algorithm, the possible clusterings are represented as medoids (w.r.t. the given metric) of variable length. The number of clusters is not required as a parameter, the method is able to find an optimal choice by means of evolutionary operators and a proper fitness function. An experimentation proves the feasibility of our method and its effectiveness in terms of clustering validity indices. With a supervised learning phase, each cluster can be assigned with a refined or newly constructed intensional definition expressed in the adopted language.
INTRODUCTION

In the context of the Semantic Web (henceforth SW) there is an extreme need of automatizing those activities which are more burdensome for the knowledge engineer, such as ontology construction, matching, and evolution. These phases can be assisted by specific learning methods, such as instance-based learning (and analogical reasoning) (Aarts, Korst, & Michiels, 2005), case-based reasoning (d’Aquin, Lieber, & Napoli, 2005), inductive generalization (Esposito, Fanizzi, Iannone, Palmisano, & Semeraro, 2004; Iannone, Palmisano, & Fanizzi, 2007; Lehmann & Hitzler, 2008) and unsupervised learning (clustering) (Fanizzi, Iannone, Palmisano, & Semeraro, 2004; Kietz & Morik, 1994), crafted for knowledge bases expressed in description logics (DLs) The Description Logic Handbook, 2003 (that are the standard representations of the field) and complying with their semantics.

In this work, we investigate on unsupervised learning for DL knowledge bases. In particular, we focus on the problem of conceptual clustering of semantically annotated resources surveying recent work on metric induction and evolutionary methods (Fanizzi, d’Amato, & Esposito, 2007). Besides, we propose the exploitation of clustering in order to detect the evolution of the ontologies over time by detecting concept drift (Widmer & Kubat, 2006) or novelties (Spinosa, Ponce de Leon Ferreira de Carvalho, & Gama, 2007) arising from the newly acquired individuals and their related assertions. Indeed, these two phenomena are mainly due to the introduction of new (previously unknown) assertions of individuals as instances of one or more concepts.

The benefits of conceptual clustering (Stepp & Michalski, 1986) in the context of semantically annotated knowledge bases are manifold. Clustering annotated resources enables the definition of new emerging concepts (concept formation) on the grounds of the concepts defined in a knowledge base; supervised methods can exploit these clusters to induce new concept definitions or to refining existing ones (ontology evolution); intensionally defined groupings may speed-up the task of search and discovery (Aarts et al., 2005; d’Amato, Staab, Fanizzi, & Esposito, 2007); a clustering may also suggest criteria for ranking the retrieved resources based on the distance from the centers.

Essentially, most of the clustering methods are based on the application of similarity (or density) measures defined over a fixed set of attributes of the domain objects. Classes of objects are taken as collections that exhibit low interclass similarity (density) and high intraclass similarity (density). These methods are rarely able to take into account some form of background knowledge that could characterize object configurations by means of global concepts and semantic relationships. This hinders the interpretation of the outcomes of these methods which is crucial in the SW perspective that enforces sharing and reusing the produced knowledge in order to enable forms of semantic interoperability across different knowledge bases and applications.

Conceptual clustering methods can answer these requirements since they have been specifically crafted for defining groups of objects through (simple) descriptions based on selected attributes (Stepp & Michalski, 1986). In the perspective, the expressiveness of the language adopted for describing objects and clusters (concepts) is extremely important. Related approaches, specifically designed for DLs representations, have recently been introduced (Fanizzi et al., 2004; Kietz & Morik, 1994). They pursue logic-based methods for attacking the problem of clustering with respect to some specific DL languages. The main drawback of these methods is that they are language-dependent, which prevents them from scaling to the standard SW representations that are mapped on complex DLs. Moreover, purely logic methods can hardly handle noisy data.

These problems motivate the investigation on similarity-based clustering methods which can be more noise-tolerant and language-independent.