PROCLAIM: An Unsupervised Approach to Discover Domain-Specific Attribute Matchings from Heterogeneous Sources

Molood Arman[1,2], Sylvain Wlodarczyk[1], Nacéra Bennacer Seghouani[2], and Francesca Bugiotti[2]

[1] Services Pétroliers Schlumberger, 34000, Montpellier, France
[2] Université Paris-Saclay, CNRS, Laboratoire de Recherche en Informatique, 91405, Orsay, France
{marman2,swlodarczyk}@slb.com
{nacera.seghouani,francesca.bugiotti}@lri.fr

Abstract. Schema matching is a critical problem in many applications where the main goal is to match attributes coming from heterogeneous sources. In this paper, we propose PROCLAIM (PROfile-based Cluster-Labeling for AttrIbute Matching), an automatic, unsupervised clustering-based approach to match attributes of a large number of heterogeneous sources. We define the concept of attribute profile to characterize the main properties of an attribute using: (i) the statistical distribution and the dimension of the attribute's values, (ii) the name and textual descriptions related to the attribute. The attribute matchings produced by PROCLAIM give the best representation of heterogeneous sources thanks to the cluster-labeling function we defined. We evaluate PROCLAIM on 45,000 different data sources coming from oil and gas authority open data website[3]. The results we obtain are promising and validate our approach.

1 Introduction

During the last years, the availability of multiple and heterogeneous data sources has given new perspectives to the schema matching problem which is a fundamental step for data integration. A large number of research works exist in the literature, the main task in these approaches is to identify the correlation between the attributes using dataset values, semantic and syntactic rules to detect the correspondence between attributes during the schema matching process [1]. Most of the works on schema integration assumed a global (mediated) schema and then tried to find a solution for better matching on mostly a pairwise matching between the source schema and the mediated schema. In this context it is very difficult to define a global schema that matches all the attributes of a given domain [10]. Moreover, real world data is always noisy and for most of integration methods, data cleaning is needed. However, in terms of big data, data cleaning is

[3] The data is published under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)