Data Mining based approach for authors disambiguation in large citation networks

Djamel Abdelkader Zighed*
ISH CNRS University of Lyon, Lyon, France - djamel@zighed.com

Adrian Tanasescu
ISH CNRS University of Lyon, Lyon, France - adrian.tanasescu@ish-lyon.cnrs.fr

Francisco Rodríguez Drumond
ISH CNRS University of Lyon, Lyon, France - frandres@gmail.com

Fabien Rico
ERIC University of Lyon, Lyon, France - fabien.rico@univ-lyon1.fr

Entity resolution is a demanding research problem for which many approaches have been proposed. Disambiguating authors of scientific publications is an instance of this problem: when facing a corpus of publications we often wish to know which articles belong to which real life authors. This task is not always possible based only on the information present in an article citation, which usually includes authors, title, year and journal. Two types of problem arise: synonymy, which corresponds to the case where the name of an individual can be written in more than one way and homonymy which describes the case where two or more individuals have the same name. In this paper we introduce a technique for addressing this task in large datasets by clustering papers based on proposed similarity measures. One can expect that the obtained clusters can be mapped to real life authors under the assumption that authors tend to write similar papers throughout time. The similarity measures we propose include a mix of string-similarity measures between relevant available fields and a novel graph-based similarity metric. The similarities produced are used to cluster papers with hierarchical agglomerative clustering. We then use a corpus of manually disambiguated papers of diverse scientific disciplines as ground truth to evaluate the performance of our technique. Our contributions include: 1) the use of a name similarity mask to determine candidate papers to be clustered, 2) a novel similarity function using a combination of string similarities as well as graph similarity metrics, and 3) a useful technique for the visualization of these similarity metrics to determine the best features for the clustering task.

Keywords: Disambiguation; Entity Resolution; Data Mining; Machine Learning.