An approach for instance based schema matching with google similarity and regular expression

ABSTRACT

Instance based schema matching is the process of comparing instances from different heterogeneous data sources in determining the correspondences of schema attributes. It is a substitutional choice when schema information is not available or might be available but worthless to be used for matching purpose. Different strategies have been used by various instance based schema matching approaches for discovering correspondences between schema attributes. These strategies are neural network, machine learning, information theoretic discrepancy and rule based. Most of these approaches treated instances including instances with numeric values as strings which prevents discovering common patterns or performing statistical computation between the numeric instances. As a consequence, this causes unidentified matches especially for numeric instances. In this paper, we propose an approach that addresses the above limitation of the previous approaches. Since we only fully exploit the instances of the schemas for this task, we rely on strategies that combine the strength of Google as a web semantic and regular expression as pattern recognition. The results show that our approach is able to find 1-1 schema matches with high accuracy in the range of 93%-99% in terms of Precision (P), Recall (R), and F-measure (F). Furthermore, the results showed that our proposed approach outperformed the previous approaches although only a sample of instances is used instead of considering the whole instances during the process of instance based schema matching as used in the previous works.

Keyword: Schema matching; Instance based schema matching; Google similarity; Regular expression