Instance Coreference Resolution in Multi-ontology Linked Data Resources

Aynaz Taheri and Mehrnoush Shamsfard

Computer Engineering Department, Shahid Beheshti University, Tehran, Iran
ay.taheri@mail.sbu.ac.ir, m-shams@sbu.ac.ir

Abstract. Web of linked data is one of the main principles for realization of semantic web ideals. In recent years, different data providers have produced many data sources in the Linking Open Data (LOD) cloud upon different schemas. Isolated published linked data sources are not themselves so beneficial for intelligent applications and agents in the context of semantic web. It is not possible to take advantage of the linked data potential capacity without integrating various data sources. The challenge of integration is not limited to instances; rather, schema heterogeneity affects discovering instances with the same identity. In this paper we propose a novel approach, SBUEI, for instance co-reference resolution between various linked data sources even with heterogeneous schemas. For this purpose, SBUEI considers the entity co-reference resolution problem in both schema and instance levels. The process of matching is applied in both levels consecutively to let the system discover identical instances. SBUEI also applies a new approach for consolidation of linked data in instance level. After finding identical instances, SBUEI searches locally around them in order to find more instances that are equal. Experiments show that SBUEI obtains promising results with high precision and recall.

Keywords: Linked Data, Coreference Resolution, Ontology, Schema, Instance, Matching.

1 Introduction

Linked data is a new trend in the semantic web context. Nowadays increasing the amount of linked data in Linking Open Data project is not the only challenge of publishing linked data; rather, matching and linking the linked data resources are also equally important and can improve the effective consuming of linked data resources. Linked data integration is one of the main challenges that become more important considering development of linked data. Without these links, we confront with isolated islands of datasets. The fourth rule of publishing linked data in [2] explains the necessity of linking URIs to each other. Therefore, extension of datasets without interlinking them is against the Linked Data principles.

In the web of linked data with so large scale, there are obviously many different schemas in the various linked data sources. Considering that there is no compulsion for data providers in utilizing specific schema, we confront with the problem of

H. Takeda et al. (Eds.): JIST 2012, LNCS 7774, pp. 129–145, 2013.

schema heterogeneity in data sources. This issue is considerable in instance coreference resolution. Paying attention to schemas in linked data consolidation has many advantages. When we are going to discover instances with unique identity in two data sources, it is a complicated process to compare all the instances of two data sources in order to find equivalents. Processing all of the instances has harmful effects on execution time and needs more computing power. However, if we know about schema matching of two data sources, it is not necessary to look up all the instances. Rather, it is enough to search only instances of two matched concepts of schemas, so performance would become better. In addition, ignoring the schema may cause precision decrease in instance matching. In many cases, the internal structure and properties of instances do not have enough information to distinguish distinct instances and this increases the possibility of wrong recognition of co-referent instances that are apparently similar in some properties in spite of their different identities.

Although ontology/schema matching can be beneficial for instance matching, it could be detrimental if it is done inefficiently. In [18] effects of ontological mismatches on data integration (in instance level) are described. They divide all types of mismatches into two groups: conceptual mismatches and explication mismatches. They represent that these kinds of mismatches such as conceptual mismatches could be harmful for instance matching and could decrease the amount of precision by wrong matching of concepts in the schema level. They do ontology matching at the first step and instance matching at the second step. Because of this sequential process, the errors of the first step can propagate into the next step.

In this paper, we propose a solution, SBUEI, to deal with the problem of instance matching and schema matching in linked data consolidation. SBUEI proposes an interleaving of instance and schema matching steps to find coreferences or unique identities in two linked data sources. SBUEI, unlike systems such as [13, 21, 27] -which uses just instance matching- or systems such as [15, 24] -which use just schema matching- exploits both levels of instance and schema matching. The main difference between SBUEI and other systems like [19], which exploit both levels, is that SBUEI exploits an interleaving of them while [19] exploits them sequentially one after the other (starts instance matching after completing schema matching). SBUEI utilizes schema matching results in instance matching and use the instance matching results in order to direct matching in schema level. SBUEI also has a new approach for instance matching.

This paper is structured as follows: section 2 discusses some related work. Section3 explains the instance coreference resolution algorithm at the first phase, and section 4 describes the schema matching algorithm at the second phase. Section 5 demonstrates the experimental results of evaluating SBUEI. Finally, section 6 concludes this paper.

2 Related Work

We divide related works in the area of entity coreference resolution in the context of semantic web into four groups: