Chapter 74
Improving the Quality of Linked Data Using Statistical Distributions

Heiko Paulheim
University of Mannheim, Germany

Christian Bizer
University of Mannheim, Germany

ABSTRACT

Linked Data on the Web is either created from structured data sources (such as relational databases), from semi-structured sources (such as Wikipedia), or from unstructured sources (such as text). In the latter two cases, the generated Linked Data will likely be noisy and incomplete. In this paper, we present two algorithms that exploit statistical distributions of properties and types for enhancing the quality of incomplete and noisy Linked Data sets: SDType adds missing type statements, and SDValidate identifies faulty statements. Neither of the algorithms uses external knowledge, i.e., they operate only on the data itself. We evaluate the algorithms on the DBpedia and NELL knowledge bases, showing that they are both accurate as well as scalable. Both algorithms have been used for building the DBpedia 3.9 release: With SDType, 3.4 million missing type statements have been added, while using SDValidate, 13,000 erroneous RDF statements have been removed from the knowledge base.

INTRODUCTION

Many of the data sets that are published as Linked Data on the Web (Bizer et al. 2009a) have been created from structured sources such as relational databases and are thus *strongly structured* (Bizer and Cyganiak, 2006). In addition, Linked Data is also extracted from *semi-structured sources*, such as Wikipedia (Bizer et al. 2009b, Lehmann et al. 2014) or from *unstructured sources* such as free text (Ramakrishnan et al. 2006, Augenstein et al. 2012, Gerber and Ngonga Ngomo 2012).

DOI: 10.4018/978-1-5225-5191-1.ch074

Linked Data which has been extracted from semi- and unstructured sources is likely to contain noise in the form of wrong RDF statements (Dutta et al., 2014). The data is also likely to be rather incomplete with respect to its schema. For example, Linked Data sets which have been generated from relational databases usually contain type information for each resource since such information is present in most databases. This does not hold for data sets that were extracted from semi-structured and unstructured sources, where that information may be missing in the original source, or the information extraction system was not able to extract it. Furthermore, heuristically extracted data sets are more likely to contain a certain level of noise in the form of wrong statements.

In order to improve the quality of such noisy and incomplete Linked Data sets, this article proposes the *SDType* method for adding missing type information to a data set, as well as the *SDValidate* method for identifying possibly wrong statements which were generated by the information extraction system. Both methods do not use any external knowledge, i.e., they exploit solely the data set itself. The proposed methods rely on statistical distributions of types and properties, i.e., characteristic distributions of the types of a property's subjects and objects. We show that both algorithms have a high accuracy and scale to large knowledge bases such as DBpedia and NELL. Both algorithms have been used to improve the quality of the DBpedia 3.9 release: With SDType, we have added 3.4 million missing type statements, while with SDValidate, 13,000 wrong statements have been identified and removed.

The rest of the article is structured as follows. We first discuss data quality issues that are particular to Linked Data sets which have been extracted from semi- and unstructured sources, i.e., noisy and incomplete data sets, and describe the two datasets used for evaluation in this article, i.e., DBpedia, and a Linked Data version of NELL. Then, we introduce the idea of using statistical distributions of types and properties for quality improvement, which underlies both algorithms discussed in this article, and we further define the two the algorithms SDType and SDValidate. For both algorithms, we discuss the evaluation on the two datasets, and describe how they have been deployed for the DBpedia release 3.9. We present an implementation of the algorithms in a relational database system, and, based on that implementation, discuss their complexity. We conclude the paper with a review of related work, a summary, and an outlook on future work.

Parts of the work presented in this article have been published as part of the conference paper "Type Inference on Noisy RDF Data" (Paulheim and Bizer, 2013). While that conference paper only discusses the SDType algorithm, this article extends the conference paper by introducing the complementary SDValidate algorithm including its evaluation, and further evaluates the differences between SDType and type inference using classical ontology reasoning. Furthermore, it compares the results achieved with both algorithms on an additional dataset, i.e., a Linked Data version of NELL.

DATA QUALITY ISSUES WITH NOISY AND INCOMPLETE LINKED DATA SETS

Data quality is not a single measure, but has multiple dimensions. Pipino et al. (2002) list several of those dimensions, ranging from accessibility to completeness. In addition, many of those dimensions cannot be assessed in a context-free manner, but depend on the task at hand, such as relevance. Thus, data quality is generally conceived as "fitness for use" (Wang et al., 1996), i.e., the capability of data to fit the requirements of a specific user given a certain use case.

Linked data sets created from semi-structured or unstructured sources face certain data quality problems that are unique to that class of data sets. The first difference is concerned with the *completeness*