Chapter 2
Towards a More Scalable Schema Matching: A Novel Approach

Sana Sellami
LIRIS, National Institute of Applied Science of Lyon, France

Aïcha-Nabila Benharkat
LIRIS, National Institute of Applied Science of Lyon, France

Youssef Amghar
LIRIS, National Institute of Applied Science of Lyon, France

ABSTRACT

With the development and the use of a large variety of DB schemas and ontologies, in many domains (e.g. semantic web, digital libraries, life science, etc), matching techniques are called to overcome the challenge of aligning and reconciling these different interrelated representations. Matching field is becoming a very attractive research topic. In this chapter, the authors are interested in studying scalable matching problem. The authors survey the approaches and tools of large scale matching, when a large number of schemas/ontologies and attributes are involved. They attempt to cover a variety of techniques for schema matching called Pair-wise and Holistic. One can acknowledge that this domain is on top of effervescence and scalable matching needs many more advances. Therefore, they propose our scalable schema matching methodology that deals with the creation of a hybrid approach combining these techniques. Their architecture includes a pre-matching approach based on XML schemas decomposition. As shown by their experiments, their proposed methodology has been evaluated and implementing in a PLASMA (Platform for LArge Scale MAtrching) prototype.

DOI: 10.4018/978-1-4666-0906-8.ch002
1 INTRODUCTION

Nowadays, the Information Technology domains (semantic web, deep web, e-business, digital libraries, life science, biology, etc) abound with a large variety of DB schemas, XML schemas or ontologies stored in many heterogeneous databases and information sources. One can observe commonly in e-business applications for example schemas with several thousand elements and expressed in different formats. Thereby, a hard problem has been brought up: solving the semantic heterogeneity in the large and perform the integration of such heterogeneous collections of schemas and ontologies. Matching techniques are solutions to automatically find correspondences between these schemas/ontologies in order to allow their integration in information systems. More precisely, matching is an operation that takes as input (e.g XML schemas, ontologies, relational database schemas) and returns the semantic similarity values of their elements. Even if matching has found considerable interest in both research and practice “in the small”, it still represents a laborious process “in the large”. The standard approaches trying to match the complete input schemas often leads to shading off performance. Various schema matching systems have been developed to solve the problem semi-automatically. Since schema matching is a semi-automatic task, efficient implementations are required to support interactive user feedback. In this context, scalable matching becomes a hard problem to be solved.

A number of approaches and principles (Rahm & Bernstein, 2001, Shvaiko & Euzenat, 2005, Do & all, 2002) have been developed for matching small or medium schemas and ontologies (50-100 components), whereas in practice, real world schemas/ontologies are voluminous (hundred or thousand components). In consequence, matching algorithms are facing up to more complicated contexts. As a result, many problems can appear, for example: performance decreasing when the matching algorithms deal with large schemas/ontologies, their complexity becomes consequently exponential, increasing human effort and poor quality of matching results is observed.

In this context, a major challenge that is still largely to be tackled is to scale up semantic matching according to two facets: a large number of schemas to be aligned or matched and very large schemas. While the former is primarily addressed in the database area, the latter has been addressed by researchers in schema and ontology matching. Based on this observation, we propose a new scalable methodology for schema matching. Our methodology supports ii) a hybrid approach trying to address the two facets based on the combination of pair-wise and holistic strategies and is deployed in three phases (pre-matching, matching and post-matching; ii) a decomposition strategy to divide large, XML schemas into small ones using tree mining technique. Our methodology has been evaluated and implementing in PLASMA (Platform for LArge Scale MAting) prototype specifically developed to this aim.

The chapter is organized as follows. Section 2 reviews related works. In section 3, we describe in detail our methodology for scalable schema matching. Section 4 presents experimental evaluation results. Finally, we conclude and discuss future works.

2 RELATED WORKS

In this section, we discuss the proposed solutions in the literature of the large scale matching problems. This issue has been tackled in holistic and pair-wise matching approaches, using different strategies e.g fragmentation, clustering, statistical, etc. We describe in the following section these different strategies and review the scalable matching tools.

2.1 Pair-Wise Matching

Being a central process for several research topics like data integration, data transformation, schema
Related Content

Energy Network Operation in the Supercomputing Era
Tianxing Cai and Neha Gupta (2015). Research and Applications in Global Supercomputing (pp. 242-262).
www.igi-global.com/chapter/energy-network-operation-in-the-supercomputing-era/124345?camid=4v1a

Towards Efficient Bounds on Completion Time and Resource Provisioning for Scheduling Workflows on Heterogeneous Processing Systems
www.igi-global.com/article/towards-efficient-bounds-on-completion-time-and-resource-provisioning-for-scheduling-workflows-on-heterogeneous-processing-systems/185774?camid=4v1a

Modeling and Analyzing of Research Topic Evolution Associated with Social Networks of Researchers
www.igi-global.com/article/modeling-and-analyzing-of-research-topic-evolution-associated-with-social-networks-of-researchers/159098?camid=4v1a

Replica Management in Data Intensive Distributed Science Applications
Ann Chervenak and Robert Schuler (2012). Data Intensive Distributed Computing: Challenges and Solutions for Large-scale Information Management (pp. 188-205).
www.igi-global.com/chapter/replica-management-data-intensive-distributed/62828?camid=4v1a