Chapter 9
Large Scale Matching Issues and Advances

Sana Sellami
LIRIS, France

Aicha-Nabila Benharkat
LIRIS, France

Youssef Amghar
LIRIS, France

ABSTRACT

Nowadays, the Information technology domains (Semantic Web, E-business, digital libraries, life science, etc) abound with a large variety of data (e.g. DB schemas, XML schemas, ontologies) and bring up a hard problem: the semantic heterogeneity. Matching techniques are called to overcome this challenge and attempts to align these data. In this chapter, the authors are interested in studying large scale matching approaches. They survey the techniques 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, as well as a set of useful optimization techniques. They compare the different existing schema/ontology matching tools. One can acknowledge that this domain is on top of effervescence and large scale matching needs many more advances. Then the authors provide conclusions concerning important open issues and potential synergies of the technologies presented.

INTRODUCTION

Recently, we are witnessing an explosive growth of data in the business and scientific area. In fact, there are many databases and information sources available through the web covering different domains: semantic Web, deep Web, e-business, biology, digital libraries, etc. In such domains, the data generated are heterogeneous and voluminous e.g schemas with several thousand elements are common in e-business applications. Currently, the greatest challenge to take up is to perform the integration of such heterogeneous collections of data. Matching techniques are solutions to automatically find correspondences between these data in order to allow their integration in information systems. Matching has found considerable interest in both research and practice. In fact, matching is an opera-

DOI: 10.4018/978-1-61520-859-3.ch009

tion that takes data as input (e.g XML schemas, ontologies, relational database schemas) and returns the semantic similarity values of their elements. However, matching these data at large scale represents a laborious process. The standard approach trying to match the complete input schemas will often lead 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 problem to be solved.

This chapter describes new research works of large scale schema and ontology matching. In the past years there has been quite an amount of research in the area of matching both for database schemas and more recently for ontologies. Several surveys (Rahm& Bernstein, 2002, Shvaiko & Euzenat, 2005) have been proposed covering many of the existing approaches. The survey proposed by (Rahm& Bernstein, 2002) is devoted to a classification of schema matching approaches and a comparative review of matching systems. The survey exposed by (Shvaiko & Euzenat, 2005) presents, as well, a new classification taking into account some novel schema/ontology matching approaches. A number of approaches and principles have been developed for matching small or medium data (schemas or ontologies). A major challenge that is still largely to be tackled is to scale up semantic matching in two ways: to a large number of data to be aligned or matched and to very large data. 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 provide a survey of work in the large scale area that differs from those proposed by (Rahm& Bernstein, 2002, Shvaiko & Euzenat, 2005). We provide in our study the main features of a large scale matching. We survey, then, the existing matching approaches at large scale called holistic and Pair-wise and we show how these approaches deal with scalability problem. We discuss the several related strategies and topics of optimization techniques, machine learning algorithms, statistical algorithms, etc. We describe the existing schema/ontology matching tools in the literature and compare them. This analysis of state of the art techniques allows us to make some conclusions and observations about the existing matching approaches and systems.

This chapter is organized as follows. Section 2 presents the motivation of large scale matching problem. Section 3 discusses the large scale matching approaches and presents a classification. In section 4, we describe the large scale matching tools. Section 5 reports some future directions and section 6 concludes this chapter.

LARGE SCALE MATCHING PROBLEM

Motivating Example

To motivate the large scale matching problem, let us consider two e-business companies: Company A and B (Figure 1).

These companies try to interoperate with sharing their internal schemas. This comprises a variety of transactions, such as exchanging product information, placing purchase orders, confirming and paying orders, which are carried out by exchanging electronic documents, or messages, between these two business partners. We have then to integrate databases of these two companies. The documents of both companies are described on e-business XML schemas.

The real-life e-business applications often process the XML data that are structured according to some standardized e-business schemas of catalogs and messages, such as OAGIS[1] or XCBL[2]. Such catalogs are developed by various individual, national and public organizations. Table 1 shows some characteristics of these E-business schemas displaying the "amazing scale" about these schemas.