Chapter 16
Matching XML Documents at Structural and Conceptual Level using Subtree Patterns

Qi Hua Pan
Curtin University, Australia

Fedja Hadzic
Curtin University, Australia

Tharam S. Dillon
Curtin University, Australia

ABSTRACT

Knowledge matching is an important problem for many emerging applications in many areas including scientific knowledge management, ontology matching, e-commerce, and enterprise application integration. Matching the concepts of heterogeneous knowledge representations is very challenging due to the difficulty of taking contextual information into account and detecting complex matches. In this chapter, we describe a knowledge matching approach that uses subtree patterns to utilize structural information for matching at the conceptual and structural level. Initially, the algorithm does not take any syntactic information into account, but rather forms candidate mappings according to their structural/contextual relationships in the knowledge structures, which are then validated using online dictionaries and string similarity measures. The approach will then automatically extract the knowledge structure that is shared among all the matched knowledge representations. Experimental evaluation is performed on a number of real world XML schemas, which demonstrates the effectiveness of the proposed approach.

DOI: 10.4018/978-1-61350-356-0.ch016
Matching XML Documents at Structural and Conceptual Level using Subtree Patterns

INTRODUCTION

Matching of heterogeneous knowledge sources has become an increasingly important problem in many areas such as e-commerce, enterprise application integration, schema matching, ontology matching/merging/alignment, knowledge management related tasks and many emerging semantic web applications. For example, many web services use XML as a unified exchange format, since it provides the extensibility and language neutrality that is the key for standards-based interoperability between different software applications (Alesso & Smith, 2006; Fensel et al., 2007). The process of discovering particular web services, composing them together and associating semantics with them in order to accomplish a specific goal is an important step toward the development of ‘semantic web services’ (Alesso & Smith, 2006; Fensel et al., 2007; Paolucci, Kawamura, Payne, & Sycara, 2002). In this process, being able to detect common knowledge structures between the information presented by the services to be integrated will be a useful step toward automation. Some other common uses of knowledge matching are to produce a shared knowledge representation for one domain or to produce a knowledge representation encompassing aspects from many domains at the same time.

Most of the current techniques used for the knowledge matching problem rely on user interaction and there are only a few methods that approach the task in a close to automatic manner. The main difficulty in this process is caused by the fact that each particular community or organization may name the same concepts differently and structure their knowledge in a unique organization-specific way. The lexical ambiguity of a word or phrase increases the difficulty of knowledge matching, which as a research topic has attracted academic contribution from the 1950’s (Ide & Véronis 1998). A recent work of Tagarelli, Longo, and Greco (2009) investigates an unsupervised word sense disambiguation method to identify the semantic relationship among the concepts underlying the constituents of structure information by referring to the WordNet. The first, and the most complex, step in the knowledge matching process is finding semantically correct matches among the concepts from heterogeneous knowledge representations. This is often referred to as the problem of conceptual level matching. Once correct mappings between the concepts have been determined, one can choose a common name to represent the concepts representing the same aspect of the domain. The next step in the knowledge matching process is that of structural level matching, which is concerned in detecting the common structure in which the concepts of the domain at hand are organized in the knowledge representation considered. In this chapter we focus on both the conceptual level and the structural level knowledge matching.

We propose a unique approach to this problem based on the utilization of previously developed tree mining algorithms (Tan, Dillon, Hadzic, Feng, & Chang, 2006; Tan, Hadzic, Dillon, & Chang, 2008; Hadzic, Tan, & Dillon, 2008) to approach the problem in a fully automated manner. The work presented in this chapter is an extension of our previously developed method for concept matching presented in (Pan, Hadzic, & Dillon, 2010). The conceptual level matching problem is addressed by utilizing the substructures of the knowledge representations to first determine the similarity of the concepts based upon their position in the knowledge substructures and the structural properties of the substructures in which they occur. This will enable us to detect candidate simple and complex matches among the concepts in the domain, by taking into account the context in which they are used. These formed mappings are evaluated for validity using online dictionaries and string similarity metrics including string edit distance and sound similarity. Once the correct matches have been identified the labels of the concepts from the formed matches are updated so that the same name is used to represent the same aspect of the domain at hand. This enables us to