Chapter 9
A Fuzzy RDF Graph-Matching Method Based on Neighborhood Similarity

Guanfeng Li
Ningxia University, China

Zongmin Ma
Nanjing University of Aeronautics and Astronautics, China

ABSTRACT

With the popularity of fuzzy RDF data, identifying correspondences among these data sources is an important task. Although there are some solutions addressing this problem in classical RDF datasets, existing methods do not consider fuzzy information which is an important property existing in fuzzy RDF graphs. In this article, we apply fuzzy graph to model the fuzzy RDF datasets and propose a novel similarity-oriented RDF graph matching approach, which makes full use of the 1-hop neighbor vertex and edge label information, and takes into account the fuzzy information of a fuzzy RDF graph. Based on the neighborhood similarity, we propose a breadth-first branch-and-bound method for fuzzy RDF graph matching, which uses a state space search method and uses truncation parameters to constrain the search. This algorithm can be used to identify the matched pairs.

INTRODUCTION

Resource description framework (RDF) (Klyne & Carroll, 2006) is a standard data model recommended by the World Wide Web Consortium (W3C) to capture resource information the context of the semantic web (Berners-Lee, Hendler, & Lassila, 2001). This model represents data as sets of triples where each triple consists of three elements that are referred to as the subject, the predicate, and the object of the triple. These triples allow users to describe arbitrary things in terms of their attributes and their relationships to other things. At the same time, information is often vague or imprecise in the real-world applications. Therefore, the study of fuzzy extension of RDF models has emerged (Ma, Li, & Yan, 2018;

DOI: 10.4018/978-1-5225-8446-9.ch009

Ma, & Yan, 2018; Mazzieri, & Dragoni, 2008; Straccia, 2009). Nowadays fuzzy RDF has been widely used in a variety of real scenarios (Pivert, Slama, & Thion, 2016; Zhang, 2017).

An increasing amount of data is becoming available in RDF format. Multiple datasets are effectively published according to the linked-data principles (Zhang, Song, He, Shi, & Dong, 2012). Integrating these datasets through interlink or fusion is needed in order to assure interoperability between the resources composing them. To do this, we need to automatically discover the correspondence between these data sources in different information stores. Data matching is the process of bringing data from different data sources together and comparing them in order to find out whether they represent the same real-world object in a given domain (Dorneles, Gonçalves, & dos Santos Mello, 2011). Efficient RDF data matching becomes the technical foundation of many tasks in Semantic Web (Li, & Ma, 2018; Zou, & Özsu, 2017).

Fuzzy RDF data have a natural representation in the form of a labeled directed graph in which the vertices present the resources and values (also called literals), and edges represent semantic relationships between resources. So, RDF data matching problem has been often addressed in terms of graph matching approach. Graph matching is usually based on the graph isomorphism or homomorphism, combining a specific application environment to find similar topology graph. Some works (Carroll, 2002) has been dedicated to the search for the best match between two graphs or subgraphs. Unfortunately, the traditional graph matching algorithms based on graph isomorphic have been proved that its complexity is NP-complete (Ullmann, 1976). For this reason, some works (Costabello, 2014; Dorneles, Gonçalves, & Mello, 2011) of approximate matching based on similarity or distance metrics use a specific index structure to reduce the complexity of RDF graph matching. However, these approximate matching approaches ignored many features of RDF graph. Firstly, these approaches only take the similarity of vertices and edges into account in RDF graph, do not concern structure among the vertices and edges. Secondly, vertices in RDF graphs have incomplete information or even anonymized information (blank vertices), and the partial neighborhood information available from a source graph will be helpful to identify entities in the target graph. More importantly, these methods cannot process fuzzy information in the matching process.

Clearly, there is a need to adopt approximate similarity matching techniques to solve the above problem. There is a kind of structural similarity approaches (Melnik, Garcia-Molina, & Rahm, 2002), in which an element in source graph G_S and an element in target graph G_T are considered similar if their respective neighborhoods within G_S and G_T are similar. This kind of approaches is particularly suitable for the lack of syntactical information (blank vertices in our case) in graphs. Note that this kind of approaches concerns only the pure structural similarity, while the utilization of other evaluation methods such as string similarities is not considered. Furthermore, it does not take edges similarities into consideration, which are key properties in RDF graph. In order to avoid anonymized information of blank vertices, improve the accuracy of vertex similarity, and fully consider the fuzzy information, we propose a 1-hop neighborhood-based similarity measure. This method combines the similarity of vertex and edge labels in the 1-hop neighborhood, and consider an edge fuzzy degree as a whole semantic unit to improve the matching accuracy.

Neighborhood-based similarity measure provides a large set of correspondences. To efficiently identify the matched pairs from the results of 1-hop neighborhood similarity measure, we need a method to solve the optimal solution of a complex combinatorial optimization problem. Branch and bound (Clausen, 1999) is an algorithm design paradigm for discrete and combinatorial optimization problems, as well as mathematical optimization. A branch-and-bound algorithm consists of a systematic enumeration of candidate solutions by means of state space search: the set of candidate solutions is thought of