Information Processing and Management 56 (2019) 1–13

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

AHAB: Aligning heterogeneous knowledge bases via iterative blocking

Chen Ling*, Gu Weidong, Tian Xiaoxue, Chen Gencai

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Keywords:
Heterogeneous knowledge base
Alignment
Iterative blocking
Candidate entity pairs

ABSTRACT

With the development of information extraction, there have been an increasing number of large-scale knowledge bases available in different domains. In recent years, a great deal of approaches have been proposed for large-scale knowledge base alignment. Most of them are based on iterative matching. If a pair of entities has been aligned, their compatible neighbors are selected as candidate entity pairs. The limitation of these methods is that they discover candidate entity pairs depending on aligned relations, which cannot be used for aligning heterogeneous knowledge bases. Only few existing methods focus on aligning heterogeneous knowledge bases, which discover candidate entity pairs just for once by traditional blocking methods. However, the performance of these methods depends on blocking keys heavily, which are hard to select. In this paper, we present an approach for aligning heterogeneous knowledge bases via iterative blocking (AHAB) to improve the discovery and refinement of candidate entity pairs. AHAB iteratively utilizes different relations for blocking, and then matches block pairs based on matched entity pairs. The Cartesian product of unmatched entities in matched block pairs forms candidate entity pairs. By filtering out dissimilar candidate entity pairs, matched entity pairs will be found. The number of matched entity pairs proliferates with iterations, which in turn helps match block pairs in each iteration. Experiments on real-world heterogeneous knowledge bases demonstrate that AHAB is able to yield a competitive performance.

1. Introduction

Knowledge bases organize human knowledge in structural form and are widely used as prior knowledge in applications, e.g., information retrieval (Han, Chen, & Tian, 2018) and question answering (Yin et al., 2016). Thanks to the rapid development of information extraction, more and more knowledge bases have been published online in recent years (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008; Hoffart, Suchanek, Berberich, & Weikum, 2013; Lehmann, Isele, Jakob, Jentzsch, & Kontokostas, 2015). Although data providers are encouraged to connect their datasets into the LOD (Linked Open Data) Cloud, most datasets are not sufficiently linked to each other. According to Schmachtenberg, Bizer, and Paulheim (2014), 44% of all datasets have no links pointing to at least one other dataset. This raises challenges in Linked Data applications. In addition, the existing links are not of high quality. Since different knowledge bases use different entity terms and the number of entities increases sharply, discovering the *owl:sameAs* links (i.e., two different URIs referring to the same real-world object) between different knowledge bases is a challenging task, which is known under various names, e.g., knowledge base alignment (Lacoste-Julien et al., 2013), entity resolution (Christophides, Efthymiou, & Stefanidis, 2015), object coreference resolution (Glaser, Jaffri, & Millard, 2009; Hu & Jia, 2015), and instance matching (Castano, Ferrara, Lorusso, & Montanelli, 2008; Rong et al., 2012).

* Corresponding author.
E-mail addresses: lingchen@zju.edu.cn (L. Chen), guweidong@zju.edu.cn (W. Gu), xxtian@zju.edu.cn (X. Tian), chengc@zju.edu.cn (G. Chen).

https://doi.org/10.1016/j.ipm.2018.08.006
Received 5 January 2018; Received in revised form 4 June 2018; Accepted 26 August 2018

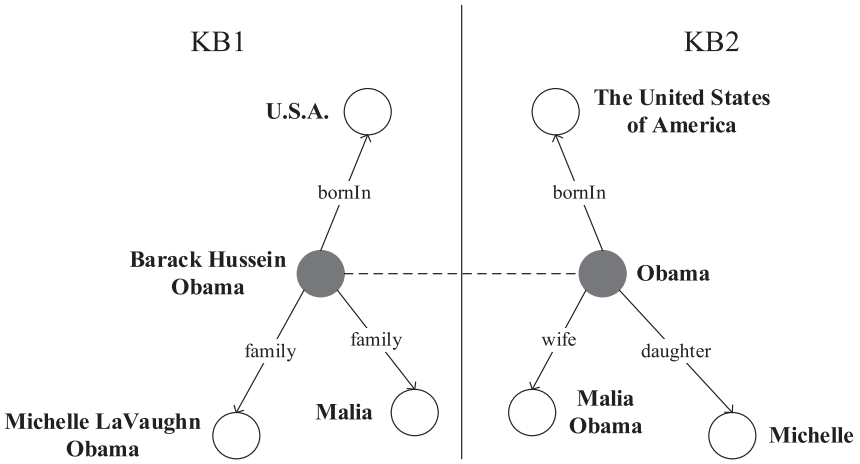

Fig.1. An example of the failure of iterative matching based methods on heterogeneous knowledge base alignment.

Considerable investments (Aswani et al., 2006; Aumueller, Do, Massmann, & Rahm, 2005; Hassell, Aleman-Meza, & Arpinar, 2006; Nagy, Vargas-Vera, & Stolarski, 2009; Niu, Rong, Zhang, & Wang, 2011) have been done to automatically discover the *owl: sameAs* links between different knowledge bases. Most of them are inefficient and cannot deal with large-scale knowledge base alignment, as they need to compare all entity pairs comprehensively. Some other methods based on iterative matching, e.g., SIGMa (Lacoste-Julien et al., 2013), SLINT (Nguyen, Ichise, & Le, 2012), and RIMOM-IM (Shao et al., 2016), are applied for large-scale knowledge base alignment. The idea of these methods is that if a pair of entities has been aligned, their compatible neighbors are selected as candidate entity pairs. Therefore, an entity is just compared with few entities from another knowledge base to improve efficiency. However, these methods mainly align homogeneous knowledge bases, i.e., two knowledge bases with lots of aligned relations and properties. They are ineffective for aligning heterogeneous knowledge bases, i.e., two knowledge bases with few aligned relations and properties, as they would miss some candidate entity pairs. An example is shown in Fig. 1. In KB1 and KB2, the entities "Barack Hussein Obama" and "Obama" have been aligned. Since the relations "bornIn" are matched, the entities "U.S.A." and "The United States of America" can be selected as a candidate entity pair. However, the relation between "Barack Hussein Obama" and "Michelle LaVaughn Obama" is described as "family" in KB1, while the relation between "Obama" and "Michelle" is described as "wife" in KB2. By the above methods, the entities "Michelle LaVaughn Obama" and "Michelle" cannot be selected as a candidate entity pair. Similarly, the entities "Malia" and "Malia Obama" also cannot be selected as a candidate entity pair.

In addition to the above methods based on the graph structure, some other works use the content of knowledge bases to discover candidate entity pairs, e.g., SERIMI (Araujo, Tran, Vries, & Schwabe, 2015). This kind of methods selects candidate entity pairs by blocking, i.e., using the literals of properties as blocking keys to distinguish entities and the discovery of candidate entity pairs does not depend on aligned relations. However, the performance of this kind of methods depends on blocking keys heavily, which are hard to select.

As mentioned above, the discovery and refinement of candidate entity pairs are two key points in knowledge base alignment research. In this paper, we propose an approach for aligning heterogeneous knowledge bases via iterative blocking (AHAB), which covers both these two key points. AHAB uses different relations for blocking in different iterations, so that more candidate entity pairs can be discovered. In each iteration, AHAB uses a divide-and-conquer strategy to discover and refine candidate entity pairs: first, for each knowledge base, select a relation randomly to divide the entities of the knowledge base into blocks, which are then matched based on matched entity pairs; next, the Cartesian product of unmatched entities in matched block pairs is treated as candidate entity pairs; finally, refine candidate entity pairs by computing their similarity. By such a design, the number of matched entity pairs proliferates with iterations, which would in turn lead to more matched block pairs.

The main contributions of this paper are summarized as follows:

(1) Propose a heterogeneous knowledge base alignment approach AHAB, which iteratively discovers and refines candidate entity pairs. Blocks are constructed within each knowledge base using different relations in different iterations, so AHAB can discover more candidate entity pairs.
(2) Present a divide-and-conquer strategy to discover and refine candidate entity pairs in each iteration, in which, based on matched entity pairs, blocks are matched first to discover candidate entity pairs, and then new matched entity pairs are discovered by refining candidate entity pairs.
(3) Evaluate the proposed approach on real-world heterogeneous knowledge bases and perform a comparison with other methods. The experimental results show that AHAB is able to yield a competitive performance.

The remainder of this paper is structured as follows. In the next section, we review the related work. Section 3 gives preliminaries and the problem definition. Section 4 introduces the framework and the details of AHAB. Section 5 reports the experimental results. Section 6 concludes and discusses future research directions.