Theoretical Computer Science 549 (2014) 101–126

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

A theoretical framework for knowledge-based entity resolution

Klaus-Dieter Schewe[a,1], Qing Wang[b,*]

[a] Software Competence Center Hagenberg and Johannes-Kepler-University Linz, Austria
[b] Research School of Computer Science, The Australian National University, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 10 June 2012
Received in revised form 1 October 2013
Accepted 1 October 2013
Available online 27 June 2014
Communicated by W. Fan

Keywords:
Entity resolution
Data matching
Record linkage
Query evaluation
Query containment
Knowledge representation
Knowledge optimization

ABSTRACT

Entity resolution is the process of determining whether a collection of entity representations refer to the same entity in the real world. In this paper we introduce a theoretical framework that supports knowledge-based entity resolution. From a logical point of view, the expressive power of the framework is equivalent to a decidable fragment of first-order logic including conjunction, disjunction and a certain form of negation. Although the framework is expressive for representing knowledge about entity resolution in a collective way, the questions that arise are: (1) how efficiently can knowledge patterns be processed; (2) how effectively can redundancy among knowledge patterns be eliminated. In answering these questions, we first study the evaluation problem for knowledge patterns. Our results show that this problem is NP-complete w.r.t. combined complexity but in PTIME w.r.t. data complexity. This nice property leads us to investigate the containment problem for knowledge patterns, which turns out to be NP-complete. We further develop a notion of optimality for knowledge patterns and a mechanism of optimizing a knowledge model (i.e. a finite set of knowledge patterns). We prove that the optimality decision problem for knowledge patterns is still NP-complete.

1. Introduction

Entity resolution is one of the major impediments affecting data quality provided by information systems. The difficulty of this problem has been widely acknowledged by research communities [10,16,21] and industry practitioners [1,37,38]. State-of-the-art approaches to entity resolution favor similarity-based methods [14]. Numerous classification techniques have been developed under a variety of perspectives such as probabilistic [21,26], cost-based [42], ruled-based [15], supervised [23], active learning [5,30,39], and collective classifications [10,16]. A common rationale behind similarity-based methods is that, the more similar two entities are, the more likely they refer to the same real-world object. However, since entities that look similar may refer to different objects, and conversely entities that look different may refer to the same objects, similarity-based methods are far from perfect. Such problems become more evident when the information about entities is inadequate. Imagine that for two entities e_1 and e_2 it is only known that they have the same name, how can we decide whether or not e_1 and e_2 refer to the same real-world object? Example 1.1 shows that, as resolving entities based on similarity is

[*] Corresponding author.
E-mail addresses: kd.schewe@scch.at, kd.schewe@faw.at (K.-D. Schewe), qing.wang@anu.edu.au (Q. Wang).

[1] The research reported in this paper was supported by the European Fund for Regional Development (DAABAA_00121) as well as the State of Upper Austria for the project *Vertical Model Integration* within the program "Regionale Wettbewerbsfähigkeit Oberösterreich 2007–2013" (Wi-219205).

http://dx.doi.org/10.1016/j.tcs.2014.06.030
0304-3975/© 2014 Elsevier B.V. All rights reserved.

AID	NAME	AFFILIATION	EMAIL
1	Q. Wang	PBRF Office, University of Otago	qing.wang@otago.ac.nz
2	Qing Wang	Dept. of Information Science, University of Otago	qwang@infoscience.otago.ac.nz
3	Qing Wang	RSCS, Australian National University	qing.wang@anu.edu.au
4	Q. Wang	CAU Kiel, Germany	wang@is.informatik.uni-kiel.de
5	Q. Wang	Dept. of Information Systems, Massey University	q.q.wang@massey.ac.nz

Fig. 1. Sample records in AUTHOR.

AUTHORGROUP$_1$
{2,3}
{1,4,5}

(a)

AUTHORGROUP$_2$
{1}
{2}
{3}
{4}
{5}

(b)

AUTHORGROUP$_2$
{1,2,4}
{3}
{5}

(c)

Fig. 2. Identifying authors.

hardly ever completely accurate, we should care for the possibility to revise decisions on entity resolution whenever more knowledge becomes available.

Example 1.1. Consider the relation AUTHOR in Fig. 1 and the question "which of these authors refer to the same person in the real world". To answer this question, traditional similarity-based methods would use certain techniques to measure the similarity between authors in AUTHOR and then group them based on their similarity. For instance, authors may be grouped based on the similarity of their names as shown in Fig. 2(a). Since authors with similar names could be different persons, and on the other hand authors with different names could be the same person, we do not actually know whether the result presented in Fig. 2(a) is accurate.

Suppose that over time we gradually acquire the following knowledge from other sources:

($K1$) *Qing Wang worked at both the PBRF Office and the Department of Information Science, University of Otago*;
($K2$) *Q. Wang studied at the CAU Kiel, Germany before joining the Department of Information Science, University of Otago*.

By $K1$, we know that the authors 1 and 2 refer to the same person. Similarly, by $K2$ we know that the authors 2 and 4 refer to the same person. Hence, the result presented in Fig. 2(a) is incorrect. Considering that the main reason for this problem is that Qing Wang is a very common name used in Asian countries, we revise the previous name-similarity-based approach by excluding all authors named "Qing Wang" or "Q. Wang". This change would yield one distinct group for each author named "Qing Wang" or "Q. Wang" by default, as shown in Fig. 2(b). These authors can only be grouped together if more specific knowledge about their entity resolution become available. In this case, by using the knowledge patterns that capture $K1$ and $K2$ (they will be presented as P_{A1} and P_{A2} in Example 2.1), the authors 1, 2 and 4 can be grouped together, while the authors 3 and 5 are still left in different groups if no more knowledge is available yet. Fig. 2(c) presents this result.

Although we can use knowledge about entity resolution to improve the results of traditional similarity-based methods, knowledge artifacts acquired from different sources at different times are often not consistent. For instance, we may acquire the following $K3$ later on, which is however conflicting with $K1$ and $K2$.

($K3$) *Q. Wang at the PBRF Office of the University of Otago has never studied or worked in Germany*.

Hence, questions concerning inconsistencies of knowledge artifacts naturally arise, such as "how can we efficiently detect the inconsistency among $K1$, $K2$ and $K3$?" and "can we reverse the decision on resolving the authors 2 and 4 if $K2$ is incorrect?". To answer such questions, we first need to represent knowledge about entity resolution in a structural and efficient way, and then build automatic reasoning tools for checking consistency of knowledge artifacts.

It is also worth noting that knowledge artifacts are often acquired at different levels of abstraction. For instance, the following $K4$ is more general than $K3$.

($K4$) *Nobody at the PBRF Office of the University of Otago has ever studied or worked in Germany*.

The aim of this paper is to develop a theoretical framework for knowledge-based entity resolution. Such a framework can incorporate traditional similarity-based methods into knowledge patterns for improving the quality of entity resolution. As the first step, we are concerned with finding a suitable knowledge representation formalism for entity resolution with attractive computational properties. In particular, we are interested in the following issues: