Information Systems 75 (2018) 1–12

Contents lists available at ScienceDirect

Information Systems

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

Reasoning about attribute value equivalence in relational data

Fengfeng Fan[a,*], Zhanhuai Li[a], Qun Chen[a], Lei Chen[b]

[a] School of Computer Science and Engineering, Northwestern Polytechnical University, 27 West Youyi Road Xian Shaanxi, PR China
[b] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 28 August 2017
Revised 22 December 2017
Accepted 16 February 2018
Available online 21 February 2018

Keywords:
Entity resolution
Attribute value matching
Relational data
Value Correlation Analysis

ABSTRACT

In relational data, identifying the distinct attribute values that refer to the same real-world entities is an essential task for many data cleaning and mining applications (e.g., duplicate record detection and functional dependency mining). The state-of-the-art approaches for attribute value matching are mainly based on string similarity among attribute values. However, these approaches may not perform well in the cases where the specified string similarity metric is not a reliable indicator for attribute value equivalence. To alleviate such limitations, we propose a new framework for attribute value matching in relational data. Firstly, we propose a novel probabilistic approach to reason about attribute value equivalence by value correlation analysis. We also propose effective methods for probabilistic equivalence reasoning with multiple attributes. Next, we present a unified framework, which incorporates both string similarity measurement and value correlation analysis by evidential reasoning. Finally, we demonstrate the effectiveness of our framework empirically on real-world datasets. Through extensive experiments, we show that our framework outperforms the string-based approaches by considerable margins on matching accuracy and achieves the desired efficiency.

1. Introduction

The same real-world entities may have different representations within or across relational databases. Variations in representation can arise from differences in storage formats, typographical errors, aliases and abbreviations. Determining attribute value equivalence is an essential task for many relational data cleaning and mining applications [1,2]. For instance, most techniques for duplicate record detection in relational data [3,4] divide each record into fields (attributes) and identify duplicate records by comparing their values on fields. Effective attribute value matching can therefore improve the accuracy of duplicate record identification. Functional dependency and conditional functional dependency mining [5,6] also requires attribute value matching to reduce noise: non-identical but equivalent attribute values could make a valid functional dependency elusive.

As pointed out by the surveys [7,8], most existing work on attribute value matching focused on reasoning about the equivalence between string data. The state-of-the-art techniques are based on measuring string similarity. A wide variety of metrics [9–12] have been proposed for this purpose. In comparison, the methods for capturing similarity in numeric data are rather primitive. Typically, the similar numbers are located by simple range queries, or treated as strings, which are then compared using string similarity metrics. Therefore, effective matching usually requires a metric to accommodate the value representation variations specific to a domain. Even though the existing string similarity metrics have been shown to be effective in various applications, they also have the fundamental limitation: a metric tuned and tested on previous problems can perform poorly on a new problem. Even though researchers have proposed adaptive algorithms [13,14] that can learn similarity metrics automatically, the difficulty of using these methods cannot be overlooked: they require significant training data and intensive human intervention. Provided with a new problem, it remains challenging to design both string similarity metric and threshold that can effectively capture the value representation variations present in the problem.

We illustrate the limitation of the string-based approach by the example as shown in Table 1. The relational records refer to research papers and each paper has four attributes, `title`, `author`, `journal`, `year`, which describe the title, authors, publication venue and publication year of the paper respectively. It can be observed that the `journal` values “Computers” and “Computer” look very similar but actually represent different publication venues. In contrast, the `journal` values “Journal on Very Large Data Bases” and “VLDB J” appear much less similar but actually refer to the same research journal. To alleviate the limitation of the string-based approach, we propose to reason about attribute value equiv-

* Corresponding author.
E-mail address: fanfengfeng@mail.nwpu.edu.cn (F. Fan).

https://doi.org/10.1016/j.is.2018.02.008
0306-4379/

Table 1
A relational table on research publications.

title	author	journal	year
Energy management in industrial plants	D. Bruneo, A. Cucinotta, A.L. Minnolo, A. Puliafito, M. Scarpa	Computers	2012
Beyond bits: the future of quantum information processing	A.M. Steane, E.G. Rieffel	Computer	2000
Priority assignment in real-time active databases	R.M. Sivasankaran, J.A. Stankovic, D. Towsley, B. Purimetla, K. Ramamritham	Journal on Very Large Data Bases	2003
A taxonomy of correctness criteria in database applications	K. Ramamritham, P.K. Chrysanthis	VLDB J	2002

alence by value correlation analysis. Note that in Table 1, the two `title` values, which are correlated with "Journal on Very Large Data Bases" and "VLDB J" respectively, have a common keyword "database", and similarly, the `author` values correlated with them share a common author "K. Ramamritham". Generally, we observe that the papers published in the same journal have a higher probability to be in the same research area than those published in different journals. Accordingly, they usually share some author names and their titles share some common keywords with higher probabilities. As a result, correlation analysis between the `journal` values and their corresponding `author` and `title` values can provide with useful clues for equivalence reasoning. More specifically, if two `journal` values are correlated with many common `author` values and many highly similar `title` values, it can be reasoned that they refer to the same journal entity with a high probability.

Note that a simple type of correlation among attribute values can be described by *functional dependency*, which specifies that the value of one attribute uniquely determines the value at another attribute. Obviously, a functional dependency can be exploited to match two attribute values. In the example shown in Table 1, suppose that each paper has a unique title. Accordingly, we have the functional dependency,

$$\mathtt{fd}_1 : \mathtt{title} \rightarrow \mathtt{journal} \tag{1}$$

As a result, two attribute values on `journal` can be determined to be equivalent if their corresponding records have the same value at the attribute `title`. Unfortunately, in practice, it is challenging to detect a clear-cut functional dependency in the presence of non-identical but equivalent attribute values, and even if it can be successfully detected, it may be of limited use in determining equivalence due to lack of matching data. Again, in the example shown in Table 1, if each record refers to a unique paper, the functional dependency, $\mathtt{fd}_1$, is then powerless in determining attribute value equivalence at `journal` because there do not exist two papers sharing a common title.

As illustrated by the motivating example, besides string similarity measurement, value correlation analysis can also be useful in reasoning about attribute value equivalence in relational data. Therefore, in this paper, we aim for a formal probabilistic model for value correlation analysis, and also a unified framework that can incorporate both string similarity measurement and value correlation analysis. Our major contributions can be summarized as follows:

1. We present a novel probabilistic approach to estimate the probability of attribute value equivalence by value correlation analysis, which reasons about the equivalence between two attribute values by analysing their correlation with other attribute values.
2. We propose a unified framework for attribute value matching in relational data. Based on both string similarity measurement and value correlation analysis, it provides a unified equivalence estimation by evidential reasoning. The proposed framework is a unified one in the sense that it can be simplified into a pure string similarity metric by setting the evidence weight of value correlation analysis to be 0.
3. We experimentally evaluate the performance of the proposed framework on real-world publicly-available datasets. Our extensive experiments show the effectiveness of the unified framework, demonstrating that it outperforms the string-based approaches by considerable margins on matching accuracy and achieves the desired efficiency.