Chapter 7.17
A Machine Learning Approach to Data Cleaning in Databases and Data Warehouses

Hamid Haidarian Shahri
University of Maryland, USA

ABSTRACT

Entity resolution (also known as duplicate elimination) is an important part of the data cleaning process, especially in data integration and warehousing, where data are gathered from distributed and inconsistent sources. Learnable string similarity measures are an active area of research in the entity resolution problem. Our proposed framework builds upon our earlier work on entity resolution, in which fuzzy rules and membership functions are defined by the user. Here, we exploit neuro-fuzzy modeling for the first time to produce a unique adaptive framework for entity resolution, which automatically learns and adapts to the specific notion of similarity at a meta-level. This framework encompasses many of the previous work on trainable and domain-specific similarity measures. Employing fuzzy inference, it removes the repetitive task of hard-coding a program based on a schema, which is usually required in previous approaches. In addition, our extensible framework is very flexible for the end user. Hence, it can be utilized in the production of an intelligent tool to increase the quality and accuracy of data.

INTRODUCTION

The problems of data quality and data cleaning are inevitable in data integration from distributed operational databases and online transaction processing (OLTP) systems (Rahm & Do, 2000). This is due to the lack of a unified set of standards spanning over all the distributed sources. One of the most challenging and resource-intensive phases of data cleaning is the removal of fuzzy duplicate records. Considering the possibility of a large number of records to be examined, the removal requires many comparisons and the comparisons demand a complex matching process.

The term *fuzzy duplicates* is used for tuples that are somehow different, but describe the same

real-world entity, that is, different syntaxes but the same semantic. Duplicate elimination (also known as entity resolution) is applicable in any database, but critical in data integration and analytical processing domains, where accurate reports and statistics are required. The data cleaning task by itself can be considered as a variant of data mining. Moreover, in data mining and knowledge discovery applications, cleaning is required before any useful knowledge can be extracted from data. Other application domains of entity resolution include data warehouses, especially for dimension tables, online analytical processing (OLAP) applications, decision support systems, on-demand (lazy) Web-based information integration systems, Web search engines, and numerous others. Therefore, an adaptive and flexible approach to detect the duplicates can be utilized as a tool in many database applications.

When data are gathered form distributed sources, differences between tuples are generally caused by four categories of problems in data, namely, the data are incomplete, incorrect, incomprehensible, or inconsistent. Some examples of the discrepancies are spelling errors; abbreviations; missing fields; inconsistent formats; invalid, wrong, or unknown codes; word transposition; and so forth as demonstrated using sample tuples in Table 1.

Very interestingly, the causes of discrepancies are quite similar to what has to be fixed in data cleaning and preprocessing in databases (Rahm & Do, 2000). For example, in the extraction, transformation, and load (ETL) process of a data warehouse, it is essential to detect and fix these problems in dirty data. That is exactly why the elimination of fuzzy duplicates should be performed as one of the last stages of the data cleaning process. In fact, for effective execution of the duplicate elimination phase, it is vital to perform a cleaning stage beforehand. In data integration, many stages of the cleaning can be implemented on the fly (for example, in a data warehouse as the data is being transferred in the ETL process). However, duplicate elimination must be performed after all those stages. That is what makes duplicate elimination distinctive from the rest of the data cleaning process (for example, change of formats, units, and so forth).

In order to detect the duplicates, the tuples have to be compared to determine their similarity. Uncertainty and ambiguity are inherent in the process

Table 1. Examples of various discrepancies in database tuples

Discrepancy Problem	Name	Address	Phone Number	ID Number	Gender
	John Dow	Lucent Laboratories	615 5544	553066	Male
Spelling Errors	John Doe	Lucent Laboratories	615 5544	553066	Male
Abbreviations	J. Dow	Lucent Lab.	615 5544	553066	Male
Missing Fields	John Dow	-	615 5544	-	Male
Inconsistent Formats	John Dow	Lucent Laboratories	(021)6155544	553066	1
Word Transposition	Dow John	Lucent Laboratories	615 5544	553066	Male