Chapter 17
Matching Attributes across Overlapping Heterogeneous Data Sources Using Mutual Information

Huimin Zhao
University of Wisconsin-Milwaukee, USA

ABSTRACT

Identifying matching attributes across heterogeneous data sources is a critical and time-consuming step in integrating the data sources. In this paper, the author proposes a method for matching the most frequently encountered types of attributes across overlapping heterogeneous data sources. The author uses mutual information as a unified measure of dependence on various types of attributes. An example is used to demonstrate the utility of the proposed method, which is useful in developing practical attribute matching tools.

INTRODUCTION

As we are continually building more databases, there is accordingly a growing need for integrating these databases. The need arises due to independent development of local data islands within an organization, business mergers and acquisitions, and collaboration across business partners. Studying distributed and heterogeneous computing environments has become an important research topic for MIS researchers (March et al., 2000).

Identifying matching attributes across heterogeneous data sources is a critical and time-consuming step in integrating the data sources (Clifton et al., 1997), either physically (e.g., by consolidating local data sources into a central data

DOI: 10.4018/978-1-61350-471-0.ch017

warehouse) or logically (e.g., by constructing a federated mediating system). Despite over two decades of extensive research, schema matching still seems to largely involve ad hoc solutions (Gal, 2006). More effective techniques that can provide analysts with automated support are still in high demand. Schema matching remains an active research area, exemplified by abundant recent publications (e.g., Bonifati et al., 2008; Bozovic & Vassalos, 2008; Kang & Naughton, 2008; Rull et al., 2008; Saleem et al., 2008; Zhao & Ram, 2008).

Recently, Zhao and Soofi (2006) proposed a method that uses mutual information to explore correspondences between free-text character attributes in heterogeneous databases. This method computes a degree of similarity between two attributes and uses mutual information to measure the dependence between attribute matching and record matching. In this paper, we further extend this method and propose a comprehensive method for matching most frequently encountered types of attributes across heterogeneous data sources that share some overlapping records. We use mutual information as a unified measure of dependence on various types of attributes. We also simplify the analysis of free-text character attributes by transformations, eliminating the need for additional attribute matching functions.

The rest of the paper is organized as follows. We first present an example of heterogeneous databases. We then review some related work in the field. We then describe the proposed approach and its application in the chosen example. Finally, we conclude the paper with contributions, limitations, and potential future research directions.

ONLINE BOOKSTORE EXAMPLE

We will use an example of heterogeneous databases for illustrative purposes. In this case, there are two book catalogs extracted from the Web sites of two leading online bookstores. The catalogs have several corresponding attributes. However, most of the attribute names are not displayed on the Web sites. We manually extracted 737 and 722 records from the Web sites of the two stores, respectively. Tables 1 and 2 show some sample entries of the two catalogs. The attribute names were manually assigned to facilitate discussion, but are not used by the attribute matching method proposed in this paper and have no effect on its result. There is a common key, the ISBN, across the two catalogs. There are 702 matching records, according to the ISBN, in the two sample tables we extracted. Note that the attribute referred to as "Author" may contain multiple authors for a book. This is what we can directly observe at the Web sites of the bookstores. We do not attempt to speculate upon the actual schemas of the back-end databases hidden in the "deep Web" (He & Chang, 2006; Su et al., 2006; Wang et al., 2004). We analyze the fields displayed at the Web sites as they are.

There are various discrepancies across the two catalogs. There are spelling errors in some attributes (e.g., "Developers" vs. "Developer's" in a book title). Some author names may be shorter in one catalog than in the other (e.g., "Oracle Corp" vs. "Oracle Corporation"). The subtitle of a book may appear in one catalog but not in the other (e.g., "HTML 4 for the World Wide Web: Visual QuickStart Guide"). Some of the edition numbers are missing. Cover is coded differently (e.g., "H" vs. "Hardcover", "P" vs. "Paperback"). Publisher may be named differently (e.g., "Osborne McGraw-Hill" vs. "McGraw-Hill Professional Book Group"). Publishing date has different formats in the two catalogs. The supplemental information is described differently (e.g., "Bk&Cd Rom" vs. "BK+CD"). Prices are recorded in number of dollars in catalog 1 and number of cents in catalog 2. The sales rank and rating are conducted independently by the two bookstores and are different in general. The task is to determine the corresponding attributes based on this sample of matching records.