Chapter 7.19
An Algebraic Approach to Data Quality Metrics for Entity Resolution Over Large Datasets

John Talburt
University of Arkansas at Little Rock, USA

Richard Wang
Massachusetts Institute of Technology, USA

Kimberly Hess
CASA 20th Judicial District, USA

Emily Kuo
Massachusetts Institute of Technology, USA

ABSTRACT

This chapter introduces abstract algebra as a means of understanding and creating data quality metrics for entity resolution, the process in which records determined to represent the same real-world entity are successively located and merged. Entity resolution is a particular form of data mining that is foundational to a number of applications in both industry and government. Examples include commercial customer recognition systems and information sharing on "persons of interest" across federal intelligence agencies. Despite the importance of these applications, most of the data quality literature focuses on measuring the intrinsic quality of individual records than the quality of record grouping or integration. In this chapter, the authors describe current research into the creation and validation of quality metrics for entity resolution, primarily in the context of customer recognition systems. The approach is based on an algebraic view of the system as creating a partition of a set of entity records based on the indicative information for the entities in question. In this view, the relative quality of entity identification between two systems can be measured in terms of the similarity between the partitions they produce. The authors discuss the

difficulty of applying statistical cluster analysis to this problem when the datasets are large and propose an alternative index suitable for these situations. They also report some preliminary experimental results and outline areas and approaches to further research in this area.

INTRODUCTION

Traditionally, data quality research and practice have revolved around describing and quantifying the intrinsic quality of individual data records or rows in a database table. However as more and more organizations continue to embrace the strategies of customer relationship management (CRM), new issues are raised related to the quality of integrating or grouping records, especially as it related to the process of entity resolution.

Most current approaches to data integration quality are rooted in the evaluation of traditional data matching or duplicate detection techniques, such as precision and recall graphs (Bilenko & Mooney, 2003). However, these techniques are inadequate for modern knowledge-based entity resolution techniques where two records for the same entity may present entirely different representations, and can only be related to each other through a priori assertions provided by an independent source of associative information.

The authors propose that casting data integration problems in set theoretic terms and applying well-developed definitions and techniques from abstract algebra and statistics can lead to productive approaches for understanding and addressing these issues, especially when applied to very large datasets on the order of 10 to 100 million records or more. The chapter also describes the application of algebraic techniques for defining metrics for grouping accuracy and consistency, including measurement taken on real-world data.

BACKGROUND

Entity resolution is the process in which records determined to represent the same real-world entity are successively located and merged (Benjelloun, Garcia-Molina, Su, & Widom, 2005). It can also be viewed as a special case of heterogeneous system interoperability (Thuraisingham, 2003). The attributes that are used to determine whether records related to two entities are the same are called "indicative information." A basic problem is that the indicative information for same entity can vary from record to record, and therefore does not always provide a consistent way to represent or label the entity. Although the specific techniques used to implement a particular entity resolution system will vary, in almost all cases the end result is that the system assigns each entity a unique "token," a symbol or string of symbols that is a placeholder for the entity. Token-based entity resolution systems fall into two broad classes, based on how the tokens are created: hash tokens and equivalence class tokens.

Hash Tokens

The simplest method for associating a token with an entity is to use an algorithm to calculate or "derive" a value for the token from the primary indicative information for the entity. The derived value is called a "hash token." For example, if the indicative information for a customer were "Robert Doe, 123 Oak St.," then the underlying binary representation of this string of characters can be put through a series of rearrangements and numeric operations that might result in a string of characters like "r7H5pK2."

The use of hash tokens for entity resolution has two drawbacks: hash collisions and lack of consistency. Hash collisions occur when the hash algorithm operating on two different arguments creates the same hash token, thus creating a many-to-one mapping from indicative information to the token representations. There are a number of