Chapter 17
Strategies for Large-Scale Entity Resolution Based on Inverted Index Data Partitioning

Yinle Zhou
IBM Corporation, USA

John R. Talburt
University of Arkansas – Little Rock, USA

ABSTRACT

Inverted indexing is a commonly used technique for improving the performance of entity resolution algorithms by reducing the number of pair-wise comparisons necessary to arrive at acceptable results. This chapter describes how inverted indexing can also be used as a data partitioning strategy to perform entity resolution on large datasets in a distributed processing environment. This chapter discusses the importance of index-to-rule alignment, pre-resolution index closure, post-resolution link closure, and workflows for record-based identity capture and update, and attribute-based identity capture and update in a distributed processing environment.

BACKGROUND

Entity Resolution

Entity resolution (ER) is the process of determining whether two references to real-world objects in an information system are referring to the same object, or to different objects (Talburt, 2011). ER has long been recognized as a key data cleaning process for removing duplicate records in database systems (Naumann & Herschel, 2010), and in entity-based data integration as a way to aggregate information about the same entity across different information sources.

DOI: 10.4018/978-1-4666-4892-0.ch017

In these types of applications, the entire ER process comprises executing a set of matching rules that link together those records determined to be equivalent (duplicate), selecting one best example, called a survivor record, from each cluster of equivalent records, discarding the duplicate records, then passing the surviving records to the next process. In this role of addressing the data quality problem of redundant and duplicate data and as a precursor to data integration, ER is fundamentally in a data cleansing tool (Herzog et al, 2007). However, ER is increasingly being used in a broader context for two important reasons.

The first is that as information quality has matured and follows more of a product management focus organizations are giving more attention to problem of not only achieving high-levels of information quality, but also sustaining information quality over time (Wang, 1998). This is evidenced by several important developments of recent years including the recognition of Sustaining Information Quality as one of the six domains in the framework of information quality developed by the International Association for Information and Data Quality (Yonke et al, 2012) as the basis for the Information Quality Certified Professional (IQCP) credential, the recent approval of the ISO 8000-110:2009 standard for master data quality, and the growing interest by organizations in adopting and investing in master data management (MDM).

Master data in an organization are the data items that reference the entities that are the organization's critical, non-fungible assets, such as customers, employees, products, and equipment. MDM comprises the policies, procedures, and infrastructure needed to accurately capture, integrate, and manage master data (Loshin 2008). MDM is essentially an effort to maintain the constraint of entity identity integrity over master data.

Entity identity integrity is one of the basic tenets of data quality that applies to the representation of a given domain of real-world entities in an information system (Maydanchik, 2007). Entity identity integrity has also been described as proper representation (Huang et al, 1999). Entity identity integrity requires that:

- Each real-world entity in the domain has one and only one representation in the information system;
- Distinct real-world entities have distinct representations in the information system.

Entity Identity Information Management

Entity Identity Information Management (EIIM) is the collection and management of identity information with the goal of sustaining entity identity integrity over time (Zhou & Talburt, 2011). It is an iterative process that combines ER and data structures representing entity identity into specific operational configurations (EIIM configurations) that when executed in concert, work to maintain the entity identity integrity of master data over time.

One of the primary motivations for introducing the EIIM model is to broaden the thinking and research about ER in two ways. The first is to bring more focus on the problem of identity management by describing different strategies and data structures for capturing and managing identity information and some of the trade-offs among these strategies. The second motivation is to place more emphasis on the temporal aspects of ER and to recognize that identity information has a life cycle that can be managed over time through the use of EIIM Configurations.

A distinguishing feature of the EIIM model is the entity identity structure (EIS), a data structure that represents the identity of a specific entity and that persists from process to process. In the model presented here, the EIS is an explicitly defined structure that exists and is maintained independently of the records being processed by the system. Although all ER systems address the issue of identity representation in some way, it is