SPECIAL ISSUE ON ENTITY RESOLUTION

Overview: The Criticality of Entity Resolution in Data and Information Quality

The topic of entity resolution (ER) has always been a key component of data and information quality. It is not surprising that Lee et al. [2006] point out “multiple sources of the same information” as one of the key root conditions of data quality problems. In particular, it is often the case that these sources reference entities such as customers, products, or accounts and that some type of ER process is required to referee their equivalence. Interest in entity resolution research and practice appears to be growing commensurate with the increase in data volume and complexity. In addition, ER applications have broadened from their roots in census counting and customer data integration to more complicated applications in areas such as health care, law enforcement, product and supply chain management, social network analytics, and financial services. The need to more accurately determine equivalence among master data references and maintain entity identity integrity in information systems has never been more critical than it is now.

Even with so much attention being focused on so called “big data” and “data analytics”, ER demands more critical attention than ever before. For example, in the commonly used map-reduce approach, the reduce process is only as effective as its ability to collate similar map instances. Conversely, the advances in rapid indexing of large datasets, brought about in the Not Only SQL (NoSQL) movement, have opened up new approaches to ER implementation that were not previously practical.

Although not comprehensive of the many areas and new directions of ER research, it is hoped that this special issue will stimulate more work in this critical aspect of data and information quality. This special issue includes articles in four areas of current ER research: ER in ontology-based systems, ER in relationship resolution, probabilistic matching and handling of uncertain decisions, and ER education.

The Semantic Web as well as many non-Web information system applications, such as referent tracking [Hogan et al. 2011], are moving from traditional relational database designs to more sophisticated ontology models in which the same entity may have many roles and relationships with other entities. The first article, “Domain-Independent Entity Coreference for Linking Ontology Instances” by Dezhao Song and Jeff Heflin, is illustrative of how ER can be applied in these contexts.

In addition to the Semantic Web, the growth of social media has drawn increasing attention to ER in the context of entity relationships that can be visualized and analyzed as a graph where the nodes represent entities and the edges represent relationships between the entities. The second article, “Adaptive Connection Strength Models for Relationship-Based Entity Resolution” by Rabia Nuray-Turan, Dmitri Kalashnikov, and Sharad Mehrotra, describes an adaptive version of the connection strength approach to relationship resolution that can be optimized for a specific domain of application and minimizes the participation of domain experts.

In a similar vein, the need to reduce the clerical review of ambiguous matches has been a focus of research since the introduction of the Fellegi-Sunter Model of Record Linkage [Fellegi and Sunter 1969]. The third article, “Indeterministic Handling of Uncertain Decisions in Deduplication” by Fabian Panse, Maurice Van Keulen, and Norbert Ritter, proposes an indeterministic approach that minimizes the false

decisions in ambiguous circumstances and the participation of domain experts in clerical review.

The fourth article, “A Graduate-Level Course on Entity Resolution and Information Quality: A Step toward ER Education” by Yinle Zhou, Eric Nelson, Fumiko Kobayashi, and John Talburt, discusses some lessons learned in offering a course in ER principles and theory as part of the Information Quality Graduate Program at the University of Arkansas at Little Rock. A great resource for ER education that was not available at the time this article was written is the recently published book on Data Matching by Peter Christen [2012].

ACKNOWLEDGMENTS

It has been a great privilege to assist in assembling this special issue of the *ACM Journal of Data and Information Quality* on entity resolution (ER). I am especially grateful to the authors who have contributed the four articles included in this edition and to the reviewers for their diligence and helpful insights. In addition special thanks to Yang Lee and Stuart Madnick, the journal's editors-in-chief, for their guidance and for making this special issue possible.

REFERENCES

- Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer.
- Fellegi, I. P. and Sunter, A. B. 1969. A theory for record linkage. *J. Amer. Stat. Assoc.* 64, 328, 1183–1210.
- Hogan, W. R., Garimalla, S., and Tariq, S. A. 2011. Representing the reality underlying demographic data. In *Proceedings of the International Conference on Biomedical Ontology*. 147–152.
- Lee, Y. W., Pipino, L. L., Funk, J. D., and Wang, R. Y. 2006. *Journey to Data Quality*. The MIT Press.

—John R. Talburt, University of Arkansas at Little Rock
Guest Editor