Guest Editorial for the Special Issue on Data Quality in Databases

Interdisciplinary research addressing the challenges of information quality touches a diversity of topics including economics, psychology, information systems, data mining, database technology, and many others. Venues such as this *ACM Journal for Data and Information Quality* (JDIQ), the annual International Conference on Information Quality (ICIQ), and the annual QDB workshop strive to create opportunities to showcase research that bridge these disciplines and achieve a common goal, namely, to provide information of appropriately high quality for a multitude of different use cases and applications. While information systems research concentrates on the modeling of many facets and dimensions of information quality, computer science has a focus on the algorithmic and data management aspects of analyzing data with respect to diverse quality criteria and maintaining or increasing quality.

This Special Issue on Data Quality in Databases highlights four of the latest computational and algorithmic results covering the following broad range of important data quality issues:

—the use of data mining techniques to analyze data quality;
—methods to implant information quality dimensions into a (sensor stream) database;
—an innovative method to clean data within a database;
—an approach to make use of data quality values by incorporating them into database queries.

Thus, these articles present methods to generate, store, improve, and query data quality values.

"Mining in Large Noisy Domains" by Manoranjan Dash and Aayush Singhania recognizes that, while sampling with replacement is a well-known approach to deal with large datasets, there are two inherent problems. One is that noise in the data can impact the sampling so that it is no longer a representative sample. More important, sampling cannot distinguish data and noise, and this can degrade the results of data mining. The research proposes the "Concise" methodology to overcome these problems. The approach is shown to work well for the main data mining tasks of classification, clustering, and association rule mining. This article makes important theoretical contributions, and, at the same time, presents the results of an extensive experimental evaluation on a variety of datasets and tasks. Thus, it has the potential to lead to the development of practical tools.

"Optimal Stopping: A Record Linkage-Approach" by George Moustakides and Vassilis Verykios revisits the record-linkage problem; it is one of the earliest data quality-oriented problems of data management that still poses unsolved challenges. The record-linkage problem tries to determine if two (or more) data entries or objects refer to the same real-world entity in the absence of some unique identifying key value. The solution consists of two steps: A search step to generate pairs of potential matches, and a matching step to determine if there is a match. Because this problem can be computationally expensive, the authors of this article introduce a method to reduce the computational time without sacrificing the accuracy or quality of the match. They do so by determining the minimum number of field comparisons that are needed to make a decision about a match. Their approach relies on results from the optimal stopping theory for Markov processes. While this article primarily makes theoretical contributions, it can lead to efficient solutions for decision making in environments where many millions of records need to be processed regularly.

"Representing Data Quality in Sensor Data Streaming Environments" by Anja Klein and Wolfgang Lehner addresses quality problems for sensor data managed in data streams. Such data is inherently error-prone, both due to faulty data creation at the sensor itself and to data processing, which must trade off data volume and efficiency. Because data cleaning is not an option in such a setting, the authors present an approach that fully discloses data quality values to the data consumer and thus allows more informed business decisions. To this end, data quality values must be captured, propagated, and aggregated through the often complex data stream process. The authors choose a set of five quality dimensions that are of particular relevance to sensor data, namely, accuracy, confidence, completeness, volume, and timeliness. Based on a classification of operators in the CQL stream-processing language, the operators are extended to not only process the data, but also to appropriately aggregate and propagate associated quality value along the query execution plan. A control technique allows an automatic trade-off between execution overhead and quality data granularity based on the "interestingness" of the underlying data.

"Incorporating Domain-Specific Information Quality Constraints into Database Queries" presents the results of a collaboration by Suzanne Embury, Paolo Missier, Sandra Sampaio, R. Mark Greenwood, and Alun Preece to use domain expertise to limit the spread of polluted data. They recognize that both humans and, more importantly, programs can inadvertently consume and thereby propagate low-quality data. Their proposed solution is to capture domain expertise as information quality (IQ) constraints. It is then up to the manager of information to guarantee that the result returned to the user is consistent with the IQ constraints. This approach has the advantage that the IQ constraints can be shared and can be of benefit to users who do not have domain expertise and could not independently generate these IQ constraints. While information managers (providers) support constraints, what is novel to this work is that the consumers control the use of the IQ constraints. They use a "quality view" to define a method to assess quality and to incorporate IQ constraints into a

query engine. They further demonstrate how IQ constraints over XML data can be incorporated into the XQuery language. As a practical contribution, they describe an execution model for quality-aware XQueries.

A total of 12 papers were submitted to the special issue and went through an extensive review process with up to 3 reviewers per article and up to 4 review cycles. This process identified these 4 articles that met both the standards of the journal as well as the objectives of the special issue. We look forward to future articles in this journal that continue the topic of this special issue, that is, to apply database techniques to data quality or analyze data quality in databases.

Felix Naumann
Hasso Plattner Institute (HPI), Potsdam, Germany
and
Louiqa Raschid
University of Maryland, College Park, USA
Guest Editors