Information Systems 38 (2013) 885–886

Contents lists available at SciVerse ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Editorial

Introduction to the special issue on data quality

Poor data quality in databases, data warehouses, and information systems affects every application domain. Many data processing tasks, such as information integration, data sharing, information retrieval, information extraction, and knowledge discovery require various forms of data preparation and consolidation with complex data processing techniques. These tasks usually assume that the data input contains no missing, inconsistent or incorrect values. This leaves a large gap between the available "dirty" data and the machinery to effectively process the data for the application purposes. In addition, tasks such as data integration and information extraction may themselves introduce errors in the data.

Many systems have been developed to address the common issues that degrade the quality of data, such as tools to capture misspellings and differences in representation for the same entity. Additional work has been done to formalize and build prototypes in a more general and principled way, for instance by defining constraints to enforce data quality rules. Some of the approaches for efficient identification and correction of data quality problems are starting to become a mature technology. For example, data cleaning operators are already embedded as first class citizens into some commercial databases.

We solicited papers on issues arising in detecting data anomalies as well as assessing, monitoring, improving, and maintaining the quality of information. We received twenty-two submissions. After two rounds of rigorous peer review and editorial review, we selected four papers for inclusion in the special issue. Each paper was reviewed by at least two expert reviewers plus the three guest editors. We believe that the acceptance rate reflects the high standards of research achieved by the accepted articles, which give an excellent overview of the most important topics in data quality research today. The significant number of high quality submissions also shows the great interest from the data management community in issues related to data quality.

The first paper, "Cross-lingual Entity Matching and Infobox Alignment in Wikipedia", by Daniel Rinser, Dustin Lange, and Felix Naumann, tackles the problem of aligning information across different language versions of the same article on Wikipedia. In fact, different versions evolve independently, so integrating their information is useful for different data quality tasks which are valuable to automatize, such as to detect and resolve inconsistencies, or to augment a page in one language with data from other language versions. To achieve this goal, the authors present a graph-based approach to identify articles in different languages representing the same real-world entity by using (and eventually correcting) the inter-language links in Wikipedia. They introduce a novel instance-based schema matching technique, which exploits information overlap across different language editions, and a robust similarity measure that can reliably quantify the similarity of strings with mixed types of data. A qualitative evaluation on manually labeled pages in four large Wikipedia editions demonstrates the effectiveness of the proposed approach.

The paper "MFIBlocks: An Effective Blocking Algorithm for Entity Resolution", by Batya Kenig and Avigdor Gal, addresses the performance of entity resolution, that is, the process of discovering tuples that correspond to the same real-world entity. The goal is to improve the execution time for blocking-based techniques. Blocking is a well known technique that separates tuples into blocks that are likely to contain matching pairs, such that only tuples within a block need to be compared. Usually, tuples are assigned to blocks using some blocking key. However, finding the "right" blocks to ensure that the effectiveness is not compromised while improving efficiency is a major challenge, and high expertise is needed in existing algorithms to construct such a key. To overcome this challenging task, the authors introduce a blocking approach that avoids selecting a blocking key altogether, relieving the user from this task. The approach is based on early evaluation of block quality based on the overall commonality of its members, which are found by determining maximal frequent item sets. An experimental evaluation on real-world and artificial data validates the approach.

In their paper on "Scheduling Strategies for Efficient ETL Execution", Anastasios Karagiannis, Panos Vassiliadis, and Alkis Simitsis discuss scheduling strategies for Extract-Transform-Load (ETL) workflows. ETL tools, which play a key role in the data quality realm, are the *de facto* standard

0306-4379/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.is.2013.03.001

for populating enterprise data warehouses with information gathered from a large variety of heterogeneous data sources. Several modules help the users to build complex flows of transformations, which run under strict performance requirements and need optimization for satisfying business objectives. The authors deal with the problem of scheduling the execution of ETL activities, such as transformations and data quality operations, with the goal of minimizing execution time and allocated memory. They investigate the effects of four scheduling policies on different flow structures and experimentally show that the use of different scheduling policies may improve ETL performance in terms of memory consumption and execution time.

The last paper is a survey by Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios titled "A taxonomy of privacy-preserving record linkage techniques". Record linkage often involves privacy-sensitive data, such as names, addresses and dates of birth; when databases are linked across organizations. The issue of how to protect the privacy and confidentiality of such sensitive information is therefore crucial. The authors present a taxonomy of fifteen dimensions, grouped into five topics, namely, privacy aspects, linkage technology, theoretical complexity, evaluation, and practical aspects, to categorize twenty-five privacy preserving record linkage techniques that have been proposed in the literature in the past fifteen years. Shortcomings of current techniques and future research directions are also highlighted in the work.

We would like to take this opportunity to thank the authors and the reviewers who contributed to this special issue for the time and effort they have put in the process. We hope readers will find this article collection useful and enjoyable as much as we have.

Mourad Ouzzani, Paolo Papotti
Qatar Computing Research Institute, Qatar Foundation, Doha, Qatar
E-mail address: mouzzani@qf.org.qa (M. Ouzzani)
ppapotti@qf.org.qa (P. Papotti)

Erhard Rahm
University of Leipzig, Leipzig, Germany
E-mail address: rahm@informatik.uni-leipzig.de

Available online 14 March 2013