Report on the First International VLDB Workshop on Clean Databases (CleanDB 2006)

Dongwon Lee
Penn State University, USA
dongwon@psu.edu

Chen Li
UC Irvine, USA
chenli@ics.uci.edu

Abstract

In this report, we provide a summary[1] of the First Int'l VLDB Workshop on Clean Databases (CleanDB 2006), which took place at Seoul, Korea, on September 11, 2006, in conjunction with the 32nd Int'l Conference on Very Large Data Bases (VLDB).

1 Workshop Overview

We proposed to organize the CleanDB workshop [6] as a forum focusing on the issues to maintain and improve the "Quality of Data" (QoD) toward clean databases. The existence of poor or erroneous data in databases causes the so-called "Garbage-in, Garbage-out" problem. For any mission-critical analysis or applications, the first and foremost task to do is to improve the quality of data. However, as the sources of data become diverse, their formats become heterogeneous, and the volume of data grows rapidly, maintaining and improving the quality of such data gets harder. To address these challenging issues, the CleanDB workshop solicited papers on database-centric data quality problems and solutions.

The program committee consisted of 27 international members. Each of 21 submissions that we have received was reviewed by at least two program committee members. The workshop finally accepted 7 full papers and 2 short papers to be presented in a one-day program. The covered topics include XML object identification, quality measures, sensor-data cleaning, or data cleaning. In addition, the program also included an invited talk by Divesh Srivastava from AT&T Labs – Research, USA.

2 Technical Program

The accepted papers were divided into three technical sessions. The papers can be downloaded from the workshop website [3], which has also additional information about the program.

2.1 Session 1

The paper titled "*Structure Aware XML Object Identification*" [9] studied the object identification problem for XML data, which is particularly hard due to the structural flexibility of XML. Motivated by the limitations of tree edit distances for approximate comparisons among XML trees, the authors defined a new distance for XML data, called the *structure aware XML distance*. The authors developed a polynomial-time algorithm to calculate the distance between XML trees. They also reported their experimental results that prove the effectiveness and efficiency of the new distance and the algorithm.

The paper titled "*QUEST: QUery-driven Exploration of Semistructured Data with ConflicTs and Partial Knowledge*" [10] studied a research challenge in integrating scientific data: data may often be missing, partially specified, or conflicting. The authors presented an assertion-based data model that captures both *value-based* and *structure-based* "nulls" in data. They introduced the `QUEST` system, which leverages the proposed model for query-driven exploration of semistructured data with conflicts and partial knowledge. The proposed approach to integration lied in enabling researchers to observe and resolve conflicts in the data by considering the context provided by the data requirements of a given research question. The authors discussed how path compatibility can be leveraged, within the context of a query, to develop a high-level understanding of conflicts and nulls in data.

The paper titled "*Column Heterogeneity as a Measure of Data Quality*" [4] studied the problem of data

[1] The summary here are taken and modified from abstracts or conclusions of actual papers.

quality. The authors identified and focused attention on a novel measure, *column heterogeneity*, that seeks to quantify the data quality problems that can arise when merging data from different sources. They identified desiderata that a column heterogeneity measure should intuitively satisfy, and discussed a promising direction of research to quantify database column heterogeneity based on using a novel combination of cluster entropy and soft clustering. They also presented a few preliminary experimental results, using diverse data sets of semantically different types, to demonstrate that this approach appears to provide a robust mechanism for identifying and quantifying database column heterogeneity.

2.2 Session 2

The paper titled "*Generic Entity Resolution with Data Confidences*" [8] considered the Entity Resolution (ER) problem, in which records determined to represent the same real-world entity are successively located and merged. The authors proposed a generic approach to the ER problem, in the sense that the functions for comparing and merging records are viewed as black-boxes. In this context, managing numerical confidences along with the data makes the ER problem more challenging to define (e.g., how should confidences of merged records be combined?), and more expensive to compute. The authors proposed a sound and flexible model for the ER problem with confidences, and propose efficient algorithms to solve it. They validated the algorithms through experiments that showed significant performance improvements over naive schemes.

The paper titled "*Circumventing Data Quality Problems Using Multiple Join Paths*" [5] proposed the Multiple Join Path (MJP) framework for obtaining high quality information by linking fields across multiple databases, when the underlying databases have poor quality data, which are characterized by violations of integrity constraints like keys and functional dependencies within and across databases. MJP associates quality scores with candidate answers by first scoring individual data paths between a pair of field values taking into account data quality with respect to specified integrity constraints, and then agglomerating scores across multiple data paths that serve as corroborating evidences for a candidate answer. The authors addressed the problem of finding the top-few (highest quality) answers in the MJP framework using novel techniques, and demonstrated the utility of their techniques using real data and their Virtual Integration Prototype testbed.

The paper titled "*In-network Outlier Cleaning for Data Collection in Sensor Networks*" [12] studied outliers that are very common in the environmental data monitored by a sensor network consisting of many inexpensive, low fidelity, and frequently failed sensors. The limited battery power and costly data transmission have introduced a new challenge for outlier cleaning in sensor networks: it must be done in-network to avoid spending energy on transmitting outliers. The authors proposed an in-network outlier cleaning approach, including wavelet based outlier correction and neighboring DTW (Dynamic Time Warping) distance-based outlier removal. The cleaning process is accomplished during multi-hop data forwarding process, and makes use of the neighboring relation in the hop-count based routing algorithm. The proposed approach guarantees that most of the outliers can be either corrected, or removed from further transmission within 2 hops. The authors have simulated a spatial-temporal correlated environmental area, and evaluated the outlier cleaning approach in it. The results showed that the approach can effectively clean the sensing data and reduce outlier traffic.

2.3 Session 3

The first paper, "*Efficiently Filtering RFID Data Streams*" [1], identified the problem of RFID data filtering and developed efficient methods to eliminate noise and duplicates from RFID observations: (1) for noise filtering, the authors proposed to maintain the original time order of observations in the output more efficiently; and (2) for duplicate elimination, they proposed to minimize memory requirement for history buffering. The effectiveness of the proposal was validated by a simulated experimentation. Their approach of data filtering is essential to provide clean and correct RFID data before the data can be further processed, transformed, and integrated for RFID-enabled pervasive applications.

Data cleaning may involve the acquisition, at some effort or expense, of high-quality data. Such data can serve not only to correct individual errors, but also to improve the reliability model for data sources. However, there has been little research into this latter role for acquired data. In the second paper titled "*Data Cleaning for Decision Support*" [2], the authors defined a new data cleaning model that allows a user to estimate the value of further data acquisition in the face of specific business decisions. As data is acquired, the reliability model of sources is updated using Bayesian techniques, thus aiding the user in both developing reasonable probability models for uncertain data and in improving the quality of that data.

The third paper, "*Cleansing Databases of Misspelled Proper Nouns*" [7], presented the results of a new data cleansing algorithm in two steps. First, string data is clustered by identifying the center and border of hyper-spherical clusters, and second, the clustered strings are cleansed with the most frequent string of the cluster. Clustering starts with a non-clustered string and computes the border b of the cluster. All strings within the overlap threshold b from the center of the cluster are assigned to one cluster. Experiments showed that the border detection is robust provided a sufficient sample size.

2.4 Keynote Address

This year's keynote address was given by Divesh Srivastava from AT&T Labs – Research with the title "*The Bellman Data Quality Browser*" [11]. In summary: data quality is a serious concern in complex industrial-scale databases, which often have thousands of tables and tens of thousands of columns. Commonly encountered problems include missing data, duplicates and default values in columns supposed to treated as keys, data inconsistencies (violation of functional dependencies), and poor quality join paths (lack of referential integrity). Compounding the data quality problems are incomplete and out-of-date metadata about the database and the processes used to populate the database. To effectively address such problems, researchers at AT&T have built the *Bellman* data quality browser. Bellman profiles the database and computes concise statistical summaries of the contents of the database, to identify approximate keys, frequent values of a field (often default values), joinable fields with estimates of join sizes paths, and to understand database dynamics (changes in a database over time).

3 Final Thought

The inaugural CleanDB 2006 was very successful. It proved the effectiveness of a one-long-day workshop with high-quality talks and papers, resulted in a lively and interesting discussion carried through the entire workshop.

After CleanDB 2006, the past and present organizers of IQIS and CleanDB workshops – Laure Berti, Naumann Felix, Helena Galhardas, Nick Koudas, Dongwon Lee, Chen Li, Monica Scannapieco, and Divesh Srivastava – have discussed and agreed to form a joint workshop, tentatively call "QDB" (Quality in Databases), to serve as *the* single forum for data cleaning and quality related research in the database community. Venky Ganti at Microsoft Research, USA and Felix Naumann at Universitat Potsdam, Germany will serve as the co-organizers of the First QDB workshop at VLDB in 2007.

Acknowledgments

CleanDB 2006 would have not been successful without the contributions of all the members of the organizing committee and the technical program committee. In particular, five senior members served as steering committee for the workshop: Surajit Chaudhuri at Microsoft Research, USA, Chin-Wan Chung at KAIST, Korea, Johann-Christoph Freytag at Humboldt-Universitat zu Berlin, Germany, Divesh Srivastava at AT&T Labs – Research, USA, and William E. Winkler at US Census, USA.

References

[1] Y. Bai, F. Wang, and P. Liu. Efficiently filtering rfid data streams. In *CleanDB Workshop*, pages 50–57, 2006.

[2] M. Benedikt, P. Bohannon, and G. Bruns. Data cleaning for decision support. In *CleanDB Workshop*, pages 58–62, 2006.

[3] CleanDB Workshop Web Site. http://pike.psu.edu/cleandb06/.

[4] B. T. Dai, N. Koudas, B. C. Ooi, D. Srivastava, and S. Venkatasubramanian. Column Heterogeneity as a Measure of Data Quality. In *CleanDB Workshop*, pages 21–24, 2006.

[5] Y. Kotidis, A. Marian, and D. Srivastava. Circumventing data quality problems using multiple join paths. In *CleanDB Workshop*, pages 33–40, 2006.

[6] D. Lee and C. Li, editors. *The First International VLDB Workshop on Clean Databases (CleanDB 2006), Seoul, Korea, September 11, 2006.*

[7] A. Mazeika and M. H. Bohlen. Cleansing databases of misspelled proper nouns. In *CleanDB Workshop*, pages 63–70, 2006.

[8] D. Menestrina, O. Benjelloun, and H. Garcia-Molina. Generic entity resolution with data confidences. In *CleanDB Workshop*, pages 25–32, 2006.

[9] D. Milano, M. Scannapieco, and T. Catarci. Structure aware xml object identification. In *CleanDB Workshop*, pages 5–12, 2006.

[10] Y. Qi, K. S. Candan, M. L. Sapino, and K. W. Kintigh. QUEST: QUery-driven Exploration of Semistructured Data with ConflicTs and Partial Knowledge. In *CleanDB Workshop*, pages 13–20, 2006.

[11] D. Srivastava. The bellman data quality browser. In *CleanDB Workshop*, pages 49–49, 2006.

[12] Y. Zhuang and L. Chen. In-network outlier cleaning for data collection in sensor networks. In *CleanDB Workshop*, pages 41–48, 2006.