10th International Workshop on Quality in Databases – QDB 2012 –

Xin Luna Dong
AT&T Labs-Research, USA
lunadong@research.att.com

Eduard Constantin Dragut
Purdue University, USA
edragut@purdue.edu

1. QDB GOALS

The problem of low-quality data in databases, data warehouses, and information systems significantly and indistinctly affects every application domain. Many data processing tasks (such as information integration, data sharing, information retrieval, and knowledge discovery from databases) require various forms of data preparation and consolidation with complex data processing techniques. These tasks usually assume that the data input conforms to nice data distributions, containing no missing, inconsistent or incorrect values. This leaves a large gap between the available "dirty" data and the available machinery to effectively process the data for the application purposes.

The term *data quality* denotes, in a broad sense, a set of properties of the data that indicates various types of error conditions. The Quality in Databases (QDB) workshop is focused on discussing various issues arising in detecting data anomalies and assessing, monitoring, improving, and maintaining the quality of data. The goals of QDB are to advance research in areas including, but not limited to:

- Duplicate detection, entity resolution, and entity reconciliation
- Conflict resolution and data fusion
- Data quality models and algebra
- Quality of linked data
- Cleaning extremely large data sets
- Data quality on the Web
- Privacy-preserving data quality
- Data quality benchmarks
- Data quality on novel data management architectures (cloud, streaming data, ...)
- Data scrubbing, data standardization, data cleaning techniques
- Quality-aware query languages and query processing techniques
- Quality-aware analytics solutions
- Data quality in data integration settings
- Role of metadata in quality measurement
- Data quality mining
- Quality of scientific, geographical, and multimedia databases
- Data quality assessment, measures and improvement methodologies
- Integrity constraints

2. QDB HISTORY

Data and information quality has become an increasingly important and interesting topic for the database community. Solutions to measure and improve the quality of data stored in databases are relevant for many areas, including data warehouses, data integration, scientific databases, and customer relationship management. QDB'12 builds on the established tradition of nine previous workshops on the topic, namely three successful *IQIS workshops* (SIGMOD 2004-2006), the *CleanDB workshop* (VLDB 2006), and five *QDB workshops* (2007-2011). The growing interest in the area is further exemplified by the recent inception of the *ACM Journal* on *Data and Information Quality*, the presence of dedicated and well-attended data-quality sessions at past editions of both *VLDB* and *SIGMOD*, and a special issue on *Towards Quality Data with Fusion and Cleaning* in the *IEEE Internet Computing*. The many positive feedback received from the workshop attendees makes us believe that QDB'12 matched the high quality and good submission level of its predecessors and attracted many participants.

3. REVIEW PROCESS

The program committee consisted of 17 renowned researchers from many different organizations. All papers received three reviews. The discussion phase was quite active and led us to finally accept 7 papers. Our selection emphasized papers on cutting-

edge research topics and promising future directions in the area of data quality and data cleaning, such as mining editing rules from existing data sources, linking Wikipedia articles with related entities and cross-lingual interwiki links, and performing data aggregation in a wireless sensor network while being aware of quality of data and energy of the sensors. The proceedings are available at `www.cyber.purdue.edu/qdb2012`.

4. WORKSHOP IN ISTANBUL

The workshop took place on August 27, 2012, the day before the VLDB conference. The workshop was attended by 39 participants, who had registered specifically for QDB 2012. It was one of the most-attended workshops at VLDB 2012.

We invited two keynote speakers at the workshop. We have also arranged two panels: one at the end of the morning sessions and focused on entity resolution, and one at the end of the afternoon sessions and focused on data cleaning and repairing. Each panel was co-ordinated by one co-chair and included the keynote speaker and the paper presenters as panelists. We encouraged questions, comments, and discussions during the panels, which inspired interesting research ideas in this area.

4.1 Flash session

We started the workshop program with a 15-minute flash session for all presenters. Each presenter had the opportunity to give a 2-minute sales talk about his or her paper. We asked the speakers to submit a brief presentation (of one or two slides) beforehand. All presenters took this opportunity and were well prepared. The presenters chose various means to steer the curiosity of the audience: by analogy with well-known problems, by emphasizing the sheer size of the manipulated data, or even with a humorous take on their problems.

The flash session followed the idea in WebDB'10 [7]. It served similar purposes: introducing the speakers even if one's actual talk can be scheduled for late afternoon; giving participants a preview of the talks to come; waking up everybody with the fast pace; and ensuring that all speakers indeed show up and are present for their talk.

4.2 Morning sessions: entity resolution

Invited talk: Prof. Erhard Rahm kicked out the morning sessions on entity resolution with a talk *"Scalable Matching of Real-world Data"*. Prof. Rahm argued that despite the existence of numerous commercial tools and research prototypes, there are still significant quality, performance, and usability issues for real-world matching tasks, such as matching products from different online shops. He described a learning-based strategy for matching products. He also talked about how to improve the scalability by cloud-based entity resolution and load-balancing schemes dealing with data skew, and presented the tool *Dedoop (Deduplication with Hadoop)* for cloud-based entity resolution.

Research session I. "Performance and efficiency of entity resolution": The first session of the workshop featured three papers targeted at various aspects of entity resolution. We give a brief description of the papers.

Dynamic Record Blocking: Efficient Linking of Massive Databases in MapReduce. Bill McNeill, Hakan Kardes, Andrew Borthwick (Intelius). This paper proposes a *dynamic blocking* algorithm that automatically chooses the blocking properties at execution time to efficiently determine which pairs of records in a data set should be examined as potential duplicates without creating the same pair across blocks. It shows how to apply the technique for linking billions of records on a Hadoop cluster.

Automatic Blocking Key Selection for Duplicate Detection based on Unigram Combinations. Tobias Vogel, Felix Naumann (Hasso-Plattner-Institut). This paper proposes a supervised technique to find suitable blocking keys automatically for a data set equipped with a gold standard. It exhaustively evaluates all possible blocking-key combinations. The presenter encouraged the audience to guess the best blocking keys for a given small data set, measured the goodness of the candidate keys and compared them with the blocking keys learned by their program at the presentation.

A Learning Method for Entity Matching. Jie Chen, Cheqing Jin, Rong Zhang, Aoying Zhou (East China Normal University). This paper presents a new learning method for the selection of the proper thresholds, distance functions and rules in the rule-based method entity matching. Given an entity matching gold standard, the selection is performed so that F-measure is optimized.

Panel: The morning panel focused on entity resolution. There were many interesting ideas proposed during the panel. Here we list a few.

- *Big Data* has raised significant attention in the research community and the industry. The keynote talk mentioned scalability improvement for record linkage for big data in the cloud computing environment [9, 10]; there are

also two talks in the morning session about improving blocking, which would have the potential to enable better parallelism. However, there are many other trends of the big data that have not been addressed much for data cleaning, such as velocity and veracity. The research questions include–*How can we adapt existing entity resolution techniques to handle the higher velocity, veracity, and variety of data from a large number of data sources? Is there any opportunity presented by the big data environment that would help improve entity resolution?*

- Knowledge graphs, social networks, and linked data are widely explored recently. There is already research on collective entity resolution that leverage the inter-connection between entities for entity resolution [1, 6, 13]. The research question is–*Can we do better in benefiting entity resolution with the rich amount of information in the networks or links?*
- Information is often temporal: there are often archives of Web data and many information is associated with a time stamp. We often need to link records across different versions of data. There has been work on linking temporal information [11, 12]. The research question is–*Can we do better in linking such temporal data, such as by mining information from the semantics context and the surrounding text?*
- We often measure entity resolution results by *F-measure*, the harmony mean of *precision* (among merged pairs of records, how many indeed refer to the same real-world entity) and *recall* (among records that refer to the same real-world entity, how many are merged). However, there are cases when we emphasize one measure over the other. For example, the panelist from *Intelius.com* mentioned that when merging records referring to people, they care more about the precision; that is, it is more troublesome for merging records that refer to different real-world persons. The research question is–*How can we allow users to specify their emphasis and automatically adapt entity resolution strategies to meet the specification?*
- There has been a lot of research going on for *crowdsourcing* [5]. On the one hand, crowdsourcing can help entity resolution, such as using the crowd to fulfill the entity-resolution task [14]. On the other hand, some crowdsourcing tasks require entity resolution, such as linking answers from different workers. The research questions include–*How to realize the many opportunities presented by crowdsourcing?*

4.3 Afternoon sessions: broader topics in data cleaning

Invited talk: Dr. Ihab Ilyas opened the afternoon sessions with his talk *"Non-destructive Cleaning: Modeling and Querying Possible Data Repairs"*. In this talk, Dr. Ilyas presented his recent endeavor in probabilistic data cleaning. He mainly focused on two problems: probabilistic record linkage and modeling, and querying possible repairs of data violating functional dependency constraints. He showed how to efficiently support relational queries under this novel model and how to allow new types of queries on the set of possible repairs.

Research session II. "Data cleaning and truth discovery": This session is right after the invited talk by Dr. Ilyas. One paper is presented in this session.

A Probabilistic Model for Estimating Real-valued Truth from Conflicting Sources. Bo Zhao, Jiawei Han (University of Illinois at Urbana-Champagne). This paper discusses a data-repairing approach other than checking dependency constraints; they built a Gaussian probabilistic model that leverages collective wisdom from multiple sources and resolves conflicts from different data sources on numerical values.

Research session III. "War stories in data quality": In the last research session three more papers told more war stories about data cleaning.

Discovering Editing Rules For Data Cleaning. Thierno Diallo, Jean-Marc Petit, Sylvie Servigne (Universite Lyon - LIRIS). This paper proposes new semantics for *editing rules* and presents pattern mining techniques for discovering editing rules from existing source relations (possibly dirty) with respect to master data, which is supposed to be clean and accurate.

Cross-Lingual Data Quality for Knowledge Base Acceleration across Wikipedia Editions. Julianna Gbls-Szab (MTA SZTAKI), Natalia Prytkova, Marc Spaniol, Gerhard Weikum (Max Planck Institute for Informatics). This paper describes an approach to discover the missing links within and across the different wikipedia editions – with each edition corresponding to a different language. The discovered links include category-to-category across different editions, article-to-article and article-to-category within the same edition. The proposed

approach was implemented and evaluated against three wikipedia editions: German, French and Hungarian.

Experiments and analysis of quality- and energy-aware data aggregation approaches in WSNs. Cinzia Cappiello, Fabio Schreiber (Politecnico di Milano). This paper gives a technique for quality- and energy-aware data aggregation in sensor networks. It partitions the stream into non-overlapping windows, and, for each window, have the sensor transmit the average value as well as individual outliers. The proposed algorithm is experimentally compared against two existing algorithms.

Panel: The second panel focused on data cleaning and repairing. There were even more active discussions in this panel. Again, we highlight a few interesting problems that the panel feels the research community should pursue.

- One of the important applications for data cleaning is for scientific data. There are missing data, replications, wrong values, imprecise values, etc. The research question is–*How can we effectively apply current data cleaning techniques on scientific data and where shall we invent new techniques?*
- Data provenance was a hot topic [2, 3]. Presumably the evolution of data and the work flow information can assist us identifying dirty data and repairing the data. The research question is–*How to leverage data provenance in a principled way for data cleaning?*
- While we focus on how to clean the mess, an alternate solution is to prevent the mess. For a single data source, this would mean preventing dirty data from creeping into the database (see existing work [4]). For data integration, this would mean carefully selecting data sources for data integration and excluding those low-quality ones (see recent work [8]). The research question is–*How to prevent dirty data before the mess in various applications?*
- There have been more and more fancy visualizations for data. Visualization of quality of data can also have practical interest: just as X-ray can help doctors identify diseases, a good visualization of data can help data analyzers identify dirty data. The research question is–*How to provide a visualization of quality of data with the goal of facilitating data cleaning?*

Throughout the co-located VLDB conference we recognized many meetings between the QDB attendees. We thus believe that the workshop met its goal of fostering an environment of vivid discussions and future collaborations, which ultimately would have the potential to advance the field of data quality. We have received very positive feedback for the flash session and the two panels. We highly recommend them for other workshops.

5. ACKNOWLEDGMENTS

We would like to thank the VLDB Endowment for sponsoring one of the keynote speaker and for providing the proceedings on USB sticks to the participants. We also thank VLDB's workshop chairs Hakan Ferhatosmanoglu, James Joshi and Andreas Wombacher; as well as VLDB's general chairs Adnan Yazici and Ling Liu, and their team for their support throughout the preparation phase and the workshop in Istanbul, Turkey. We thank Microsoft's CMT team for providing the submission and reviewing platform. We thank Cyber Center at Purdue University for their support in hosting and maintaining the Web site of the workshop (`www.cyber.purdue.edu/qdb2012`). Finally, we thank our great committee members and authors of the submissions, without whom the workshop would be impossible.

6. REFERENCES

[1] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (ACM-TKDD)*, 1(1):1–36, 2007.

[2] P. Buneman and J. Cheney. Provenance in databases. In *Proc. of SIGMOD*, 2007.

[3] P. Buneman, J. Cheney, W.-C. Tan, and S. Vansummeren. Curated databases. In *Proc. of PODS*, 2008.

[4] S. Chen, X. L. Dong, L. V. Lakshmanan, and D. Srivastava. We challenge you to certify your update. In *Sigmod*, 2011.

[5] A. Doan, M. J. Franklin, D. Kossmann, and T. Kraska. Crowdsourcing applications and platforms: A data management perspective. In *VLDB*, pages 1508–1509, 2011.

[6] X. Dong, A. Y. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proc. of SIGMOD*, pages 85–96, 2005.

[7] X. L. Dong and F. Naumann. 13th international workshop on the web and databases: Webdb 2010. *SIGMOD Record*, 39(3):37–39, 2010.

[8] X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. *PVLDB*, 6, 2013.

[9] L. Kolb, A. Thor, and E. Rahm. Dedoop: Efficient deduplication with hadoop. In *VLDB*, pages 1878–1881, 2012.
[10] L. Kolb, A. Thor, and E. Rahm. Load balancing for mapreduce-based entity resolution. In *ICDE*, pages 618–629, 2012.
[11] P. Li, X. L. Dong, A. Maurino, and D. Srivastava. Linking temporal records. *PVLDB*, 4(11):956–967, 2011.
[12] P. Li, H. Wang, C. Tziviskou, X. L. Dong, X. Liu, A. Maurino, and D. Srivastava. Chronos: Facilitating history discovery by linking temporal records. In *VLDB*, 2012.
[13] P. Singla and P. Domingos. Object identification with attribute-mediated dependences. In *Ninth European Conference on Principles and Practice of Knowledge Discovery in Databaes (PKDD)*, pages 297–308, 2005.
[14] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. *PVLDB*, 5(11):1483–1494, 2012.