Report on the 8th International Workshop on Quality in Databases (QDB10)

Andrea Maurino
Department of Informatics Systems and Communication
University of Milano - Bicocca
Via Bicocca degli Arcimboldi 8
20136, Milano, Italy
maurino@disco.unimib.it

Cinzia Cappiello
Department of Electronics and Information
Politecnico di Milano
Via Ponzio 34/5
20122, Milano, Italy
cappiello@elet.polimi.it

Panos Vassiliadis
Dept. of Computer Science
University of Ioannina
Ioannina, 45110, Hellas
pvassil@cs.uoi.gr

Kai-Uwe Sattler
FG Datenbanken und Informationssysteme
Ilmenau University of Technology
Postfach 100 565
D-98684 Ilmenau
kus@tu-ilmenau.de

1. INTRODUCTION

The eighth international workshop on Quality in Database was held in Singapore, on September 13th, 2010 and co-located with the 36th Conference on Very Large DataBase (VLDB). The main objective of the workshop was to address the challenge to detect data anomalies and assess, monitor, improve, and maintain the quality of information.

The workshop attracted 12 submissions from Asia, Australia, Europe, and the United States, out of which the Program Committee finally accepted 9 full papers. The accepted papers focused on important issues especially related to Data Quality assessment, Entity Matching, and Information Overloading.

2. QUALITY IN DATABASES: OPEN ISSUES

QDB 2010 was the eighth workshop addressing the challenges of quality in databases. Significant research contributions were presented in this edition. Anyway, the existing research works in the area of data and information quality are still far from maturity and significant room for progress exists. The participants agreed that many open challenges still remain. It is possible to classify them into three general areas:

- Improvement of the comparison of algorithms and evaluation through DQ standards
- Further investigation of well known DQ issues (privacy and visualization)
- Application of DQ in new domains (such as linked open data, and Data as a Service)

Concerning the first point, discussions with participants highlighted the lack of standards for comparing different solutions, algorithms and tools. In fact, a lot of papers create their own homemade benchmarks or golden rules to demonstrate the efficiency and the effectiveness of algorithms, but it is very rare that such data are shared to support cross comparison analysis. Participants suggested to organize specific data quality events to create occasions in which data quality researchers can compare and discuss novel ideas such as an international contest about data quality evaluation similar to the KDD cup `http://www.sigkdd.org/kddcup/` or the semantic Web service challenge `http://sws-challenge.org/wiki/index.php/Main_Page`. One concrete example, related to personal data only, is the "name game" workshop series organized within the APE-INV project `http://www.academicpatenting.eu`.

As regards the second open problem, participants agreed that DQ research should focus more on visualization and privacy issues. In fact, the visualization of the results of DQ activities (ranging from assessment to record linkage and data improvement) is a crucial point for a larger dissemination of DQ researches. It includes typical problems related to the visualization of large data sets (as in data mining field), but also it needs more tailored solutions to underline errors or to interprete the results of DQ activities. In this field, mashups seem to be a promising technology for easily integrating DQ re-

sults and supporting improvement decisions. Moreover, an important aspect when integrating data from a large number of heterogeneous sources under diverse ownerships is the provenance of data or parts thereof[2]; provenance denotes the origin of data and can also include information on processing or reasoning operations carried out on the data. In addition, provenance enables effective support of trust mechanisms and policies for privacy and rights management. In the last years many solutions for provenance models and management mechanisms have been published[4]. However, according to participants a lot of problems are still opens. For example, there is the need for a trust management system for improving accountability, building on provenance, especially in the context of data aggregation.

Finally, workshop paticipants also agreed that data quality research should also focus on the definition of methods for the quality assessment and improvement of data modelled on the basis of new paradigms for information management. Linked open data, for example, is a set of principles to share in the Web environment open data [1]. This can bring a paradigmatic shift from the classical relational data integration architecture towards new Web based solutions. In this specific scenario data are retrieved from heterogeneous data sources (e.g., relational, graph and stream-like data) and there are not appropriate methods to assess, preserve and improve the quality of such data sets. In fact, while the requirements for the quality assessment of closed (mostly corporate) data sources are well understood, several open issues raise when quality assessment has to be performed in autonomous and distributed data sources where quality-related meta-information is typically sparse and the quality of the meta-information is uncertain.

Another new paradigm for data management that it is worth to consider is the Data as a Service. Until now, "data" and "services" have been always considered two different concepts characterized by different problems, approaches, models and tools. Nevertheless in the last years there is a growing interest to the Xaas approach. X as a services, where X could be software [6] or a platform [3], introduces the possibility to consider Data as a Service and consequently data can be managed by means of typical service oriented solutions. In this context, data quality could play, for example, a fundamental role in the selection of services that is commonly based on other functional and non-functional properties (e..g., response time, availability). The convergence of data and service approaches could be the first steps toward the definition of a new and holistic theory where data and service are considered as two faces of the same problem [5].

3. KEYNOTE PRESENTATION

The keynote speech, titled "SOLOMON: Seeking the Truth Via Copying Detection" was delivered by Xin Luna Dong, AT&T. In the information era, a large amount of information sources are available and easily accessible. However, freely accessible information is often unreliable: it is often accessed by data quality problems in terms of relevance, accuracy, or authority. Moreover, the information diffusion enabled by Web technologies negatively impacts on data quality issues since errors can be easily propagated. The identification of copying content between information sources could be a valuable help for the users to filter relevant data. In this keynote Xin Luna Dong presented the SOLOMON tool that supports the discovery of copying relationships between structured data source to improve data integration features. She also explained which are the research open issues for leveraging redundancy and obtain quality from Web sources. Open issues in this field are mainly related to source selection (e.g., how many sources are sufficient for aggregation?), source integration (e.g., source ordering in online query answering) and data visualization (e.g., task-driven source exploration).

4. RESEARCH PAPERS

The technical paper session consisted of nine presentations, whose main points are summarized next. Together, they give a glimpse to the exciting new developments on data and information quality.

The paper titled "Quality Assessment Social Networks: A Novel Approach for Assessing the Quality of Information on the Web" by Tomas Knap, Irena Mlynkova introduces a Web Quality Assessment model, which is a model for the ranking of Web resources on the grounds of a quality assessment (QA) score, involving profiles and policies for the management of the resources. Since people in social networks might benefit from adopting profile properties from other people with which they are linked, the paper introduces the concept of QA social networks and algorithms for the successive application of relevant QA policies (i.e., policies of trusted users) to a person's retrieved resources.

The goal of "Deriving Effectiveness Measures for Data Quality Rules" by Lei Jiang, Alex Borgida, Daniele Barone, John Mylopoulos is the evaluation of data quality rules. Broadly speaking, data quality rules detect errors and inconsistencies and they

play an important role in data quality assessment. Starting from the results of previous research, the authors propose a quantitative framework for measuring and comparing data quality rules in terms of their effectiveness. Effectiveness formulas are built from variables that represent probabilistic assumptions about the occurrence of errors in data values, and earlier work gave examples of how to derive these formulas in an ad-hoc fashion. The presented approach involves several steps, including building Bayesian network graphs, adding (symbolic) probabilities to the nodes in the graph, and deriving effectiveness formulas. The approach is implemented in Python, and the paper reports its evaluation results.

Soumaya Ben Hassine-Guetari, Jérôme Darmont, Jean-Hugues Chauchat with the paper "Aggregation of data quality metrics using the Choquet integral" present a solution that uses the Choquet integral to aggregate different data quality metrics into a single score. When comparing different (data) items, many quality dimensions might be used along with their respective metrics. When two items A and B have different scores over different dimensions, it is not straightforward to compute a global qualifying score to facilitate their comparison. In this perspective, the aggregation of data quality metrics can be the solution for computing a global and objective data quality score. The authors contribute to the solution of the problem by suggesting how the Choquet integral might provide an answer.

The paper "Data Partitioning for Parallel Entity Matching" written by Toralf Kirsten, Lars Kolb, Michael Hartung, Anika Groß, Hanna Köpcke, and Erhard Rahm deals with an important problem of entity matching, which is an important and difficult step for integrating Web data. In order to reduce the execution time for matching algorithms, the authors investigate how entity matching can be performed in parallel on a distributed infrastructure. The paper proposes different strategies to partition the input data and generate multiple match tasks that can be independently executed. One of the suggested strategies supports both blocking to reduce the search space for matching and parallel matching to improve efficiency. Special attention is given to the number and size of data partitions as they impact the overall communication overhead and memory requirements of individual match tasks. The caching of input entities and affinity-based scheduling of matching tasks is also considered. The authors also discuss the tool that they have developed in a service-based, distributed infrastructure as well as the detailed evaluation of their method.

An interesting tool for duplicate detection is proposed in "DuDe: The Duplicate Detection Toolkit" by Uwe Draisbach and Felix Naumann. Duplicate detection, also known as entity matching or record linkage, has been a research topic for several decades. The challenge is to effectively and efficiently identify pairs of records that represent the same real world entity. Researchers have developed and described a variety of methods to measure the similarity of records and/or to reduce the number of required comparisons. Comparing these methods to each other is essential to assess their quality and efficiency. However, it is still difficult to compare results, as differences can always be found in the evaluated data sets, the similarity measures, the implementation of the algorithms, or simply the hardware on which the code is executed. To face this challenge, the paper discusses the development of a comprehensive duplicate detection toolkit named DuDe. DuDe provides multiple methods and data sets for duplicate detection and consists of several components with clear interfaces that can be easily served with individual code.

An original problem is raised by Wolfgang Gottesheim, Norbert Baumgartner, Stefan Mitsch, Werner Retschitzegger, and Wieland Schwinger with the paper "Improving Situation Awareness" related to information overloading. Information overload is a severe problem for operators of large-scale control systems, as such systems typically provide a vast amount of information about a large number of real-world objects. Systems supporting situation awareness have recently gained attention as way to help operators to grasp the overall meaning of available information. To fulfill this task, data quality has to be ensured by assessment and improvement strategies. In this paper, a vision towards a methodology for data quality assessment and improvement for situation awareness systems is presented.

The management of conditional functional dependencies is an important topic described in "Extending Matching Rules with Conditions" by Shaoxu Song, Lei Chen, and Jeffrey Yu. Matching dependencies (mds) have recently been proposed in order to make dependencies tolerant to various information representations, and proved useful in data quality applications such as record matching. Instead of a strict identification function in traditional dependency syntax (e.g., functional dependencies), mds specify dependencies based on similarity matching quality. However, in practice, mds may still be too strict and only hold in a subset of tuples in a relation. Thereby, the paper proposes conditional

matching dependencies (cmds), which bind matching dependencies only in a certain part of a table. Compared to mds, cmds have more expressive power that enables them to satisfy wider application needs. The paper includes a discussion of both theoretical and practical issues of cmds, including inferring cmds, irreducible cmds with less redundancy and, the discovery of cmds from data as well as the experimental evaluation of cmd discovery algorithms.

Finally, Peter Yeh, and Colin Puri show in "Discovering Conditional Functional Dependencies to Detect Data Inconsistencies" an approach that exploits conditional functional dependencies for detecting inconsistencies in data and hence improves data quality. The approach has been empirically evaluated on three real-world data sets, and the paper discusses the performance of the proposed approach in terms of precision, recall, and runtime. Moreover, a comparison between the presented approach and an established, state-of-the-art solution shows that the presented approach outperforms this solution across the previously mentioned dimensions. Finally, the paper describes efforts to deploy the approach as part of an enterprise tool to accelerate data quality efforts such as data profiling and cleansing.

5. ACKNOWLEDGMENTS

We would like to thank the program committee members, keynote speakers, authors and attendees, for making QDB 2010 a successful workshop. Finally, we also express our great appreciation for the support from the VLDB 2010 organization.

6. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[2] P. Buneman and W. C. Tan. Provenance in databases. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, *SIGMOD Conference*, pages 1171–1173. ACM, 2007.

[3] E. Keller and J. Rexford. The "platform as a service" model for networking. In *Proceedings of the 2010 internet network management conference on Research on enterprise networking*, INM/WREN'10, pages 4–4, Berkeley, CA, USA, 2010. USENIX Association.

[4] D. L. McGuinness, J. Michaelis, and L. Moreau, editors. *Provenance and Annotation of Data and Processes - Third International Provenance and Annotation Workshop, IPAW 2010, Troy, NY, USA, June 15-16, 2010. Revised Selected Papers*, volume 6378 of *Lecture Notes in Computer Science*. Springer, 2010.

[5] M. Palmonari, A. Sala, A. Maurino, F. Guerra, G. Pasi, and G. Frisoni. Aggregated search of data and services. *Inf. Syst.*, 36(2):134–150, 2011.

[6] W. Sun, K. Zhang, S.-K. Chen, X. Zhang, and H. Liang. Software as a service: An integration perspective. In B. Kramer, K.-J. Lin, and P. Narasimhan, editors, *Service-Oriented Computing ICSOC 2007*, volume 4749 of *Lecture Notes in Computer Science*, pages 558–569. Springer Berlin / Heidelberg, 2007.