

Report from the First and Second International Workshops on Information Quality in Information Systems - IQIS 2004 and IQIS 2005 in conjunction with ACM SIGMOD/PODS Conferences

Monica Scannapieco
Università di Roma La
Sapienza, Italy
monscan@dis.uniroma1.it

Laure Berti-Équille
IRISA, University of Rennes I
France
berti@irisa.fr

ABSTRACT

This report summarizes the constructive discussions of the first two editions of the International Workshop on Information Quality in Information Systems, IQIS 2004 and IQIS 2005, held respectively in Paris, France, on June 13, 2004 and in Baltimore, MD, USA, on June 17, 2005.

1. WORKSHOP SCOPE

The problem of poor data quality stored in database-backed information systems is widespread in the governmental, commercial, and industrial environments.

Alarming situations with various information quality problems cannot be ignored anymore and, theoretical as well as pragmatic approaches are urgently needed to be proposed and validated. As a consequence, information quality is now becoming one of the hot topics of emerging interest in the academic and industrial communities.

Many processes and applications (such as information system integration, information retrieval, and knowledge discovery from databases) require various forms of data preparation or repair with several data processing and cleaning techniques, because the data input to the application-dedicated algorithms is assumed to conform 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 process the data for application-specific purposes.

The first and second editions of the International Workshop on Information Quality in Information Systems (IQIS), in conjunction with ACM SIGMOD/PODS conferences 2004 and 2005, focused on database-centric issues in data quality.

2. WORKSHOP PROGRAM

The main goal of the IQIS workshops is to be a forum for researchers, engineers, students, and practitioners from the database, knowledge discovery and data mining, information system engineering as well as statistics communities that have an in-depth interest in information quality in database-backed information systems, and also in the various techniques of data preparation, detection of inconsistent, contradictory or improbable data, data cleaning, and

in ETL systems. The participants of the two editions of the workshop were invited to cover practical and theoretical issues of data quality. In the course of each workshop, various new approaches were presented to tackle the various problems of data quality in different application domains. In addition, the limitations and shortcomings of current approaches were discussed, and possible directions for future research in the domain were outlined. Areas of interest of these one-day events included:

- Metrics for information and data quality
- Quality-aware query languages and query processing
- Quality-aware integration
- Detection of outliers, duplicates, and inconsistencies
- Entity resolution, record and data linkage
- Intelligent data preparation and data cleaning
- Models, methodologies, and frameworks for information quality
- Application-driven information quality: bioinformatics, CRM, scientific applications
- Data type-dependent information quality: Web, multimedia, XML data.

2.1 IQIS 2004 Discussions

The topics for the technical sessions of the 2004 edition (see IQIS 2004 web site at <http://www.hiqiq.de/iqis/>) were structured in three sessions covering three relevant areas of the information quality research: data transformation and duplicate detection, assessment, and quality-driven information integration.

The keynote speech “Quality-Aware Data Integration in Peer-to-Peer Systems” was given by Maurizio Lenzerini (Università di Roma La Sapienza). He emphasized the importance of information quality in peer-to-peer system [14]. In these systems, every peer acts as both client and server, and provides part of the overall information available from a distributed environment, without relying on a single global view. He identified the research issues of data integration

in peer-to-peer systems, and outlined the impact of the notion of peer quality on both the semantics of the data integration systems, and the algorithms for query processing. Interesting discussions started because recent research has highlighted the importance of data quality issues (e.g., freshness [3]) in environments characterized by extensive data replication. Query processing in P2P environments has to include a record matching activity designed to exploit the presence of multiple overlapping sets of data. Moreover, complex data transformations for solving schema and data heterogeneities have to be applied [6]. Duplicated copies of the same data have to be compared in order to select and construct a best quality copy which is then returned for the global query and also submitted to the organizations having provided low quality copies of the same data.

In the thread of the discussions, performing duplicate detection on complex data appeared to be one of the main issues of the area discussed in this workshop.

Two papers on duplicate detection were then presented: the first one regarding XML duplicate detection [22] and the second one proposed a cost optimal decision model that takes into account the cost of erroneous classifications when deciding if two records are duplicates or not [21]. The dual problem of efficiently detecting patterns of conflicts in a pair of overlapping data sources has also been advocated by [19] and [18].

As the second theme of the workshop, the implications of information quality in data integration were considered with a specific focus on dealing with inconsistent data at query processing time. In [9], the authors proposed the definition of a formal semantics in a Global-As-View (GAV) data integration system when data retrieved at sources do not satisfy constraints on the global schema. A generalized framework for query answering in presence of inconsistent data sources has also been proposed in [15]. Finally, the results concerning information quality assessment including a methodology for data quality assessment from the user perspective [5] were presented.

2.2 IQIS 2005 Discussions

The topics for the technical sessions of the 2005 edition (see IQIS 2005 web site at <http://iqis.irisa.fr>) were structured in three main sessions covering: data quality models, record linkage, and statistics and clustering for ensuring data quality.

IQIS 2005 invited two keynote speakers. The first keynote speech “Handling Data Quality in Entity Resolution” was given by Hector Garcia-Molina (Stanford University). He discussed the challenges of the entity resolution problem under uncertain data (*i.e.*, the identification problem of matching records from multiple information sources that correspond to the same real-world entity) [10]. The second keynote speech “Methods and Analyses for Determining Quality” was given by William E. Winkler (U.S. Bureau of the Census, Statistical Research)[14]. He described situations where properly chosen metrics may indicate that data quality is not sufficiently high for monitoring processes, for modeling, and for data mining. Additionally, he described generalized methods that allow a skilled individual to perform massive clean-up of files in some situations.

Following the discussions first raised by Hector Garcia-Molina on the various issues of entity resolution, the main themes of IQIS 2005 workshop were centered on data link-

age and data cleaning. Interesting exchanges between Dongwon Lee and Sharad Mehrotra started on the complementarity of their respective approaches for data linkage: the first one [13] presented a sampling-based approximate join algorithm considering the main problems commonly occurring in large-scale bibliographic digital libraries with *mixed citations* of different but homonymic scholars and *split citations* of the same author appearing under different name variants; the approach of [7] for object consolidation analyzed not only object features, but also additional information such as inter-object relationships.

Joining the discussion, the authors of [12] proposed two error-tolerant measures for identifying related attributes across independent databases to integrate based on similarities in their data.

The other important issue raised by record linkage that was discussed in the workshop concerns the privacy: in [2], the authors proposed a secure blocking scheme to improve significantly the performance of record linkage techniques while being secure.

The discussions were completed and focused on data cleaning: [8] proposed a data cleaning approach, based on modeling data dependencies with Markov networks. Belief propagation is used to compute the marginal and posterior probabilities, so as to infer missing values or to correct errors. In [11], the authors proposed a new framework for implementing active data warehousing with ETL activities over queue networks with minimal overhead and smooth upgrade of the data source systems.

In the session dedicated to data quality models, the authors of [17] proposed a model for formal data quality agreements between data providers and data consumers and they described an algorithm for dealing with constraints on the completeness of a query result with respect to a reference data source. Applied to the biological domain, [16] extended the semi-structured data model to include quality metadata with computing and updating useful data quality measures. In [20], the authors described a technique to rank data sources by characterizing data sources that agree with accurate or high-quality data sources as likely accurate.

In complex distributed systems (such as electronic payments processing systems), the authors of [4] proposed a framework that integrates three classes of models for detecting statistically significant changes from baselines, for explaining the reason of change occurrences and for preventing data quality problems. In the last session dedicated to statistics and clustering techniques, [1] presented an algorithm for clustering mixed numerical and categorical data sets with associated confidence values to represent the certainty of correctness of the categorical values. An interesting discussion initiated between the academics and practitioners (VISA, Census, AT&T, etc.) concerned the need of benchmarks and massive sets of “real and dirty” data to test and compare the approaches.

For details of the papers and keynote speeches including slides from all presentations, please visit the workshop websites. The IQIS 2005 proceedings appear in the ACM Digital Library.

3. FUTURE PLANS FOR IQIS

The approaches discussed in the workshops were mostly based on academic research. As the workshop participants agreed on this argument, to evaluate the real value of the

approaches, they need to be tested and applied on real world applications on very large data sets. It was further remarked that cooperation between academia and industry is important, if not essential, for further development and testing of the approaches with practical data benchmarks dedicated to specific data quality problems (such as record linkage or duplicate detection). The two editions of IQIS workshop achieved their goal to provide a forum for interactive discussions. On several occasions during each one-day workshop, items for a common research agenda were debated, and enthusiastic feedbacks have been collected from the participants. Plans for future collaborations were discussed. Moreover, the third edition of the IQIS workshop will be organized on June 30th in conjunction with ACM SIGMOD/PODS 2006 conference in Chicago, IL, USA (see IQIS 2006 web site at <http://queens.db.toronto.edu/iqis2006/>).

4. ACKNOWLEDGMENTS

We would like to thank the program committee members, the keynote speakers, the authors of all submitted papers, and all the workshop participants. Special thanks are given to AT&T Labs Research and to the Università di Roma La Sapienza for their support.

5. ADDITIONAL AUTHORS

Additional authors: Felix Naumann (Humboldt-Universität zu Berlin, Germany, email: naumann@informatik.hu-berlin.de), co-chair of IQIS 2004, Carlo Batini (Università di Milano 'Bicocca', Italy, email: batini@disco.unimib.it) and Divesh Srivastava (AT&T Labs Research, Florham, NJ, USA, email: divesh@research.att.com), co-chairs of IQIS 2005.

6. REFERENCES

- [1] Bill Andreopoulos, Aijun An, and Xiaogang Wang, Clustering Mixed Numerical and Low Quality Categorical Data: Significance Metrics on a Yeast Example, *IQIS 2005*.
- [2] Ali Al-Lawati, Dongwon Lee, and Patrick McDaniel, Blocking-Aware Private Record Linkage, *IQIS 2005*.
- [3] Mokrane Bouzeghoub and Verònika Peralta, A Framework for Analysis of Data Freshness, *IQIS 2004*.
- [4] Joseph Bugajski, Robert L. Grossman, Eric Sumner, and Zhao Tang, An Event Based Framework for Improving Information Quality That Integrates Baseline Models, Causal Models and Formal Reference Models, *IQIS 2005*.
- [5] Cinzia Cappelletto, Chiara Francalanci, and Barbara Pernici, Data Quality Assessment from the User's Perspective, *IQIS 2004*.
- [6] Paulo Carreira and Helena Galhardas, Execution of Data Mappers, *IQIS 2004*.
- [7] Zhaoqi Chen, Dmitri V. Kalashnikov, and Sharad Mehrotra, Exploiting Relationships for Object Consolidation, *IQIS 2005*.
- [8] Fang Chu, Yizhou Wang, D. Stott Parker, and Carlo Zaniolo, Data Cleaning Using Belief Propagation, *IQIS 2005*.
- [9] Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, and Riccardo Rosati, Tackling Inconsistencies in Data Integration through Source Preferences, *IQIS 2004*.
- [10] Hector Garcia-Molina, Handling Data Quality in Entity Resolution, *IQIS 2005*.
- [11] Alexandros Karakasidis, Panos Vassiliadis, and Evaggelia Pitoura, ETL Queues for Active Data Warehousing, *IQIS 2005*.
- [12] Andreas Koeller, and Vinay Keelara, Approximate Matching of Textual Domain Attributes for Information Source Integration, *IQIS 2005*.
- [13] Dongwon Lee, Byung-Won On, Jaewoo Kang, and Sanghyun Park, Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries, *IQIS 2005*.
- [14] Maurizio Lenzerini, Quality-Aware Data Integration in Peer-to-Peer Systems, *IQIS 2004*.
- [15] Zoran Majkic, General Framework for Query Answering in Data Quality Cooperative Information Systems, *IQIS 2004*.
- [16] Alexandra Martinez, and Joachim Hammer, Making Quality Count in Biological Data Sources, *IQIS 2005*.
- [17] Paolo Missier, and Suzanne Embury, Provider Issues in Quality-Constrained Data Provisioning, *IQIS 2005*.
- [18] Amihai Motro, Philipp Anokhin, and Aybar C. Acar, Utility-based Resolution of Data Inconsistencies, *IQIS 2004*.
- [19] Heiko Müller, Ulf Leser, and Johann-Christoph Freytag, Mining for Patterns in Contradictory Data, *IQIS 2004*.
- [20] Raymond K. Pon, and Alfonso F. Cardenas, Data Quality Inference, *IQIS 2005*.
- [21] Vassilios S. Verykios, and George V. Moustakides, A Generalized Cost Optimal Decision Model for Record Matching, *IQIS 2004*.
- [22] Melanie Weis and Felix Naumann, Detecting Duplicate Objects in XML Documents, *IQIS 2004*.
- [23] William E. Winkler, Methods and Analyses for Determining Quality, *IQIS 2005*.