Data Quality and Provenance in Information Integration

(Position Paper for Workshop on Information Integration)

Yi Chen*
Arizona State University
yi@asu.edu

1 Motivation

Data integration and data exchange have drawn much attention in recent years as they are the channel for collaboration among different groups. While the problems of schema mapping, instance matching, and update reconciliation have been extensively studied, the issue of data quality in information integration requires further investigation. It is critical for data receivers to understand the quality of the data, especially in the domain of science, engineering, federal and medical, in order to perform quality data processing and analysis.

Data of various quality levels widely exists in many applications due to several reasons, as presented in a special issue of Data Engineering [6] in March 2006. First, automated data analysis and information extraction tools are error-prone, such as observed in the Avatar system [6]. Second, data integrated from different sources can be uncertain or even conflicting, depending on the trustworthiness of its sources and the quality of the data mapping procedures [1]. Third, experimental results and sensor data are observed in conditions that may be subject to a range of variability in natural environment, mechanical defects, contaminated samples, etc, such as noted in the Data Furnace project [6]. Furthermore, human errors in generating and processing data need to be considered, such operation errors, biases, insufficient knowledge, etc. Since data cleaning is in general expensive and even impossible, it is inevitable to have data of various quality levels. Therefore, it is important for users to be aware of the data quality in order to use the data appropriately.

Besides data quality indicators, a data receiver in information integration often would like to find out how the quality indicators are assessed and what are sources that affect the quality. This type of information can be referred as *data provenance*. Though data provenance has been studied with respect to database queries [2, 5], scientific users are interested in general data provenance information, such as the software used, the experimental steps, the raw input data and parameters that are used in an experiment to produce the results, i.e., the *workflow* information [3]. Such provenance information is important for data receivers to evaluate the reliability of the data and, when necessary, revise part of the workflow to improve data quality.

2 Challenges

To measure and manage the quality of data in information integration, we propose to annotate confidence levels with the data as its quality indicators, and provide accesses to the provenance information, i.e. the workflow that produces the data. Many technical challenges need to be addressed.

First, how should we measure data quality? The assessment of data quality is typically application-dependent. For example, automated data analysis and information extraction tools can provide the confidence about the data they produce [6]. Attempts have also been made recently on building a mass collaboration infrastructure and involving a community to assess data quality, such as the CIM [6] and CbioC project (to be introduced in Section 3).

Second, given data along with its quality indicators, how should we record and propagate the indicators through query evaluation and data transformation. This demands an effective database management system that explicitly quantifies our confidence about the data as probabilities, supports efficient query evaluation, provides confidence indicators for query results, and maintains confidence measurement when new evidences are available. Toward this goal, there have been attempts to extend database systems to store and query probabilistic data [14, 13, 12, 7]. However, many technical challenges involved remain unaddressed, as observed in [6].

Furthermore, how should we design a workflow management system to record and query the provenance for data generated from experiments? In building such a system, we need to investigate techniques of recording workflow information, modeling the relationship between worksflows and the datasets they produce, and supporting user queries on datasets, workflow and across them.

*This is joint work with Dr. Chitta Baral, Susan Davidson, Graciela Gonzalez, Subbarao Kambhampati, and Zoé Lacroix.

3 Projects at Arizona State University

Several projects are being developed at Arizona State University to address the data quality issue in information integration applications.

Quality Assessment: The *CBioC* (Collaborative Bio Curation, cbioc.org) project aims at storing knowledge buried in biomedical literature in databases by leveraging information extraction, data integration and mass collaborative scientific annotations. To measure the quality of the uncertain data, CBioC provides a mass collaborative annotation infrastructure, where a large community of scientists can participate in data curation and measure data quality voluntarily. Then a social network will be built in order to provide a reliable data quality assessment based on the accuracy of automated extraction, scientist votes, voters' credit, the quality of the publication where the extraction is derived, and so on.

Quality Management: Given input data with the estimated reliability, the *QUADS* (QUality-Aware Database System) project [11] addresses the problems of storing, querying, and updating such data effectively. Driven by applications that generate uncertain data not readily amenable to a relational representation, such as information extraction and data integration, QUADS represents uncertain data using a probabilistic XML data model. Its query engine correctly propagates the confidence measurements from source data to query results. Physical data models and optimization techniques will be investigated to achieve query evaluation efficiency.

Querying Autonomous Uncertain Databases: The *QUIC* (Querying Under Imprecision and Uncertainty) project addresses the issue of querying incomplete data in autonomous databases [9, 8]. Existing query processing techniques for incomplete or probabilistic databases [14] are not appropriate for autonomous web databases since a mediator usually does not have update capabilities over the autonomous databases and cannot replace the missing values with assertions or guessed values. The goal of QUIC is to enable query evaluation on incomplete autonomous databases, such that not only sound answers but also data items with missing values but yet highly relevant to queries will be returned in a ranked fashion, without modifying the underlying data sources.

Data Provenance and Workflow: To help data receivers to understand the data and its quality, the *ProtolDB* project [10, 4] addresses the issue of recording and querying the provenance of experimental data. It differentiates workflow design that captures the aim of experiments, the implementation that specifies the resources selected to execute the each task in a workflow, and the execution of a workflow. By recording various components of a workflow and the datasets collected from the workflow execution, and building links among them, ProtolDB will allow scientists to store, compare and revise complex workflows, as well as express queries across workflows and data.

4 Directions for Future Research

We have identified several challenges in addressing the data quality issue in information integration: assessing data quality reliably; managing and propagating quality indicators appropriately through data transformation; and tracing provenance for data quality explanation. Research projects are being conducted to address these challenges, such as the ones introduced in Section 3. After tackling each individual challenge, the goal of future research is to provide a unified framework that provides a comprehensive solution to data quality management in information integration.

Acknowledgments

The ProtocolDB project is funded by NSF IIS 0612273.

References

[1] P. Andritsos, A. Fuxman, and R. J. Miller. Clean Answers over Dirty Databases: A Probabilistic Approach. In *ICDE'06*.

[2] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In *ICDT'01*.

[3] I.-M. A. Chen and V. M. Markowitz. Modeling scientific experiments with an object data model. In *ICDE '95*.

[4] S. Cohen, S. C. Boulakia, and S. B. Davidson. Towards a model of provenance and user views in scientific workflows. In *DILS'06*.

[5] Y. Cui and J. Widom. Lineage tracing for general data warehouse transformations. In *VLDB'01*.

[6] M. Garofalakis and D. Suciu. Sepcial issue on probabilistic data management. *IEEE Data Engineerng Bulletin*, 29(1), March 2006.

[7] E. Hung, L. Getoor, and V. S. Subrahmanian. PXML: A Probabilistic Semistructured Data Model and Algebra. In *ICDE'03*.

[8] S. Kambhampati, Y. Chen, J. Fan, H. Khatri, U. Nambiar, and G. Wolf, Handling Imprecison & Incompleteness in Autonomous Databases, Technical Report TR-06-014, ASU, 2006.

[9] H. Khatri, J. Fan, Y. Chen, and S. Kambhampati. Query Processing over Incomplete Autonomous Databases, Technical Report TR-06-006, ASU, 2006.

[10] N. Kwasnikowska, Y. Chen, and Z. Lacroix. Modeling and storing scientific protocols. In *KSinBIT'06*.

[11] T. Li, Q. Shao, and Y. Chen. PEPX: A Query-Friendly Probabilistic XML Database. In *CIKM'06*.

[12] A. Nierman and H. V. Jagadish. ProTDB: Probabilistic Data in XML. In *VLDB'02*.

[13] A. D. Sarma, O. Benjelloun, A. Halevy, and J. Widom. Working models for uncertain data. In *ICDE'06*.

[14] D. Suciu and N. Dalvi. Tutorial: Foundations of probabilistic answers to queries. In *SIGMOD'05*.