Affiliates Workshop on Data Quality: Challenges for Computer Science and Statistics

Abstracts

Data Quality and Reconciliation

Munir Cochinwala
Telcordia Technologies

In this talk we will give an overview of the data quality research program at Telcordia. At Telcordia we use an in-house research prototype to handle data reconciliation and data quality analysis. The prototype and accompanying methodology includes rapid generation of appropriate pre-processing and matching rules. The prototype uses a modular JavaBeans-based architecture that allows for customized matching functions and iterative runs that build upon previously learned information. Telcordia has been able to provide significant insights to clients who recognize that they have data reconciliation problems but cannot determine root causes effectively when using currently available off-the-shelf tools.

Information Quality Processes and Technologies: Information Quality in Practice

Larry P. English
Information Impact International, Inc.

It is clear that information quality is no longer irrelevant nor a luxury in information systems. It is a requirement for sustainable competitive advantage in the new economy. The Industrial Age matured when quality processes such as continuous process improvement (CPI), total quality management (TQM), and Kaizen transformed manufacturing processes, eliminating the costs of scrap and rework. We are now seeing the maturing of the Information Age as a result of applying the same quality processes to the information product-the new currency of the new economy.

This presentation presents the state-of-the-art in information quality improvement processes as applied by leading edge companies. Processes for information quality assessment and improvement are described, as are the classifications of information quality technologies, describing the strengths and limitations in information quality management. Organizations

around the world who have implemented successful information quality processes are used to illustrate techniques and cultural transformation required for a sustainable information quality environment:

- Information quality: what it is and what it isn't
- Trends in information quality processes and methods
- Classifications of and trends in information quality technologies
- Information quality improvement: the maturing of information management
- Systemic culture change requirements to sustain information quality initiatives

The Statistical Administrative Records System and Administrative Records 2000 Experiment: System Design, Successes and Challenges

Dean H. Judson
U.S. Census Bureau

The Statistical Administrative Records System (StARS) is an attempt to link six major federal administrative records files together into a composite database that will be used to simulate an administrative records census. The six major files include the IRS 1040 master file, the IRS 1099 and information returns file, the Medicare beneficiary database, the Selective Service System registrant file, the HUD-TRACS tenant rental assistance file, and the Indian Health Service file. A seventh file, the Census NUMIDENT (which is a translation of the Social Security Administration NUMIDENT file), is used as a "lookup" file for social security number (SSN) validation and as a source of demographic characteristics.

These six files are edited to create standardized person and address records, unduplicated, validated, missing fields are imputed, and finally merged to create statistical composite records. The presentation will focus on the challenges that each of these processes create, including thresholds for observation, database ontologies, unduplication and matching decisions, and the relationship between a dynamic database and a dynamic population. The presentation will conclude with a brief description of the Administrative Records Experiment (AREX 2000), which is a test of an administrative records census in two test sites in 2000.

Challenges in Improving Information Quality

Ann Thornton
Deloitte & Touche LLP

While most agree that good data is preferable to bad data, an organization that sets out to assess and/or improve information quality faces several challenges:

- Defining the importance of information quality: Assessing the costs and benefits of improvements to information quality is difficult and requires participation of those using the information as well as those contributing information.

- Assessing information quality: A daunting amount of analysis can be required to perform a detailed assessment (for example, at the data element level). Therefore, assessments of the importance of each data element are used to prioritize work effort.

- Addressing information quality problems: Performing root cause analysis and appropriate corrective actions is often preferable to short-term fixes, but process improvement requires commitment.

- Ongoing measurement and monitoring: Metrics for information quality can be difficult to construct.

Developing Data Warehouses with Quality in Mind

Yannis Vassiliou
National Technical University of Athens

Data Warehouses provide large-scale caches of historic data. They lie between information sources gained externally or through online transaction processing systems (OLTP), and decision support or data mining queries following the vision of online analytic processing (OLAP). In developing and operating Data Warehouses, one can distinguish between different processes, each of which raises quality considerations. Using the framework of a general formal architecture developed in the DWQ project, this talk discusses some of the research conclusions regarding development of Data Warehouses considering also quality factors. Basic processes and the related quality dimensions are considered. The quality factors, metrics and measurement methods are also presented.

A Decision Model for Cost Optimal Record Matching

Vassilis Verykios
Drexel University

In an error-free system with perfectly clean data, the construction of a global view of the data consists of linking or joining two or more tables on their key fields. Unfortunately, most of the time, data stored in real life database systems are neither carefully controlled for quality nor necessarily defined commonly across different data sources. As a result, the creation of such a global data view resorts to approximate joins. In this talk, an optimal solution is proposed for the matching or the linking of database record pairs in the presence of inconsistencies, errors, or missing values in the data.

Existing models for record matching rely on decision rules that minimize the probability of error, which is the probability that a pair of records is assigned to the wrong class. Often in practice, minimizing the probability of error is not the best criterion to design a decision rule because the misclassifications of different samples may have different consequences. In this talk, we present a decision model, which minimizes the mean cost of making a decision, by assigning a different cost to each kind of misclassification. More specifically, (a) we present a decision rule, (b) we prove that this rule is optimal with respect to the cost of the decision making process, and (c) we compute the probabilities of the two types of errors that incur when this rule is applied. We also present a closed form decision model for a class of comparison vectors having conditionally independent binary components and finally we demonstrate some experimental results from applying the proposed model.

Raising the Bar for Data Quality in the New Millenium

Richard Wang
Boston University

The tutorial will provide an overview of the research conducted at the MIT Total Data Quality Management (TDQM) program over the last decade emphasizing the management of information as a product. We will discuss the approach of the MIT TDQM program advocating the institutionalizing of TDQM programs for long-term benefits. Concepts such as multi-dimensional data quality, data quality metrics, evaluation of the user's assessment of data quality, and data production map will be presented in this context. Research directions stemming from recent work on *Quality Information and Knowledge* (Prentice Hall, 1999), *Data Quality* (Kluwer Academic Publisher, 2000), *Journey to Data Quality: A Roadmap to Higher Productivity* (in preparation), and *Data Quality in the Health Care Industry* (in preparation) will be touched on.

Data Quality for Large Transaction Streams

Allan R. Wilks
AT&T Labs - Research

Data analysis for very large datasets is particularly challenging when the arrival of the data is relentless -- it's like drinking from a fire hose. An example is the 50 GB/day stream of transaction call detail from the AT&T long distance network. This talk will describe data quality issues for this stream from several perspectives, including:

- Timely detection of data anomalies
- Maintaining data integrity through software and hardware integrity
- The impact of subject-matter expertise on quality of results.

Record Linkage Methods

William E. Winkler
U.S. Census Bureau

Record Linkage is used for identifying duplicates within files and merging sets of files. This talk describes methods and software. Names, addresses, and other components in a file are initially parsed into corresponding components such as first names and house numbers that are more easily compared. The model of Fellegi and Sunter generalizes recent work on Bayesian Networks. It is used to get matching or classification scores that rank the relative quality of matches. In some situations, a generalized EM algorithm can be used to obtain optimal matching parameters through unsupervised learning. In other situations relatively small amounts of training data can be combined with large amounts of unlabelled data. String comparators account for partial agreement between strings. A generalized assignment algorithm optimizes sets of matches.