New Directions for Data Quality Mining

Laure Berti-Équille
University of Rennes 1, France
berti@irisa.fr

Tamraparni Dasu
AT&T Labs-Research, NJ, USA
tamr@research.att.com

KDD 2009: Paris, France
June 28, 2009
Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study
Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits
What is Low Data Quality?

- Missing data
- Erroneous data
- Data anomalies
- Duplicates
- Inconsistent data
- Out-of-date data
- Undocumented data
Data Quality Problems

DATA

TYPE

Continuous
Discrete
Binary
Sequence
Nominal
Categorical
Multimedia

RELATIONSHIP

Structural (record)
Sequential
Graph-based
Temporal
Spatial
Spatio-Temporal

DATA QUALITY PROBLEM

TYPE

Missing data
Anomalous data
Duplicate data
Inconsistent data
Obsolete data

CARDINALITY

Single-Point
Collection

DETECTION MODE

Model-based
Data distribution-based
Constraint-based
Pattern-based

Geomedia

ACACGTGT
John Doe
0101010101

Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits
## Example 1

**Data quality problems in a relational DB**

**SIGKDD Executive Committee**

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>City, State, Zip, Country</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piatetsky-Shapiro G.,PhD</td>
<td>KDDnuggets</td>
<td>USA</td>
<td>617-264-9914</td>
</tr>
<tr>
<td>Usama Fayyad</td>
<td>Yahoo, Sunnyvale, CA</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>Jiawei Han</td>
<td>Univ. of Illinois</td>
<td>IL 61801, USA</td>
<td>(217) 333-6903</td>
</tr>
<tr>
<td>Usama Fayyad</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raghu Ramakrishnan</td>
<td>U. Wisconsin-Madison</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sonata Sarawagi</td>
<td>IIT Bombay</td>
<td>Mumbai-400076 USA</td>
<td></td>
</tr>
<tr>
<td>Robert Grossman</td>
<td>Open Data Group</td>
<td>Chicago IL, USA</td>
<td>999-999-9999</td>
</tr>
<tr>
<td>David Jensen</td>
<td>U. of Massachusetts</td>
<td>Amherst, MA, USA</td>
<td>111-111-1111</td>
</tr>
</tbody>
</table>

**Data quality problems**

- **Typos**
- **Duplicates**
- **Misfielded Value**
- **Inconsistent Value**
- **Obsolete Value**
- **Missing Value**
- **Incomplete Value**
- **Non-standard representation**
Example 2

Outliers

Rejection area: Data space excluding the area defined between 2% and 98% quantiles for X and Y

Rejection area based on:
Mahalanobis_dist(cov(X,Y)) > \chi^2(.98,2)

Legitimate outliers or data quality problems?
Example 3

Disguised missing data

Some are obvious...
Detectable with syntactical or domain constraints

Phone number: 999-999-9999

Others are not....
Could be suspected because the data distribution doesn’t conform to the expected model

Histogram of DoBs per day of the year

Histogram of online shopping customers per age category

2% patients in the obstetrical emergency service are male...
Example 4

Censored and Truncated Data

Phone call datasets

e.g., phone calls whose duration is $< 1$ second or $> 6$ hours

datasets

<table>
<thead>
<tr>
<th>D1</th>
<th>Left truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2</td>
<td>Right censored</td>
</tr>
<tr>
<td>D3</td>
<td>Left and right censored</td>
</tr>
<tr>
<td>D4</td>
<td>Left truncated and right censored</td>
</tr>
<tr>
<td>D5</td>
<td>Complete</td>
</tr>
</tbody>
</table>

Truncated - Data point is dropped if it exceeds or falls below a certain bound.

Censored - Data is bounded to a fixed min/max limit or a default value.
Example 5

Time-Dependent Anomalies: Unusual patterns in graph analysis
e.g., Fraud, Cyber intrusion, homeland security

\[\text{Normal Scan Pattern} \quad \text{Abnormal Scan Pattern} \quad \text{Abnormal Scan Pattern}\]

\[\text{High volume communications with unknown IP addresses} \quad \text{Data loss due to transmission problems}\]
Example 6

**Anomalous subsequence**
e.g., ECGs

Asystole or data quality problem?

**Deviants in time-series**
e.g., Sensors

Real phenomenon or sensor drift?
Example 7

Contradictions between Images and Text

Abuse of tags

Arbutus tree

Tags
- arbutus
- tree
- galano
- island
- amsterdam
- animal
- animals
- april
- architecture
- art

Duplicates

Fraud

EBay Listings That Sell For Dummies - Collier, Mars
Current Price: £11.72
View similar items...

EBAY Listings That Sell Dummies E-Commerce B VALUE BOOK
Current Price: £12.89
View similar items...

EBay Listings That Sell for Dummies Book | Marsha Colli
Current Price: £13.19
View similar items...

Ebay Listings that sell for Dummies On CD- Cheap Book
Current Price: £22.99
View similar items...

eBay Listings That Sell For Dummies
Current Price: £13.19
View similar items...

Guide ID: 10000000001248134 created: 29/06/06 (updated 19/02/09) from http://reviews.ebay.co.uk/Item-Picture-Description-Theft-Prevent-eBay-Fraud_W0QQugidZ10000000001248134
False information

Example 8

Steve Jobs obituary published by Bloomberg

An obituary of very-much-alive Apple founder Steve Jobs has been accidentally published by the respected Bloomberg business news wire.

By Matthew Moore
Last Updated: 7:05PM BST 28 Aug 2008

Steve Jobs was described as the man who "refashioned the mobile phone" in the erroneous obituary. Photo: REUTERS

The story, marked "Hold for release - Do not use", was sent in error to the news service's thousands of corporate clients.
Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits
What is Data Quality?

A “subtle” combination of measurable dimensions:

- **Accuracy**
  - KDD’09 location is in Paris, France

- **Consistency**
  - Only one KDD conference per year

- **Completeness**
  - Every past KDD conference had a location

- **Freshness**
  - The location of the current KDD conference is in Paris

- **Uniqueness – no duplicate**
  - KDD is a conference, not the French hip-hop rap band
  - KDD’09, KDD 2009 and Knowledge Discovery and Data mining 2009 are the same conference edition
Data Quality Research: A World of Possibilities

4 Disciplines
- Statistics
- Database
- Knowledge Engineering
- IT Process and Workflow Management

4 Types of approach
- Prevention
- Diagnostic
- Correction
- Adaptation

5 Levels
- Dimensions
- Models
- Techniques
- Tools
- Methodologies
From the DB perspective

Data Quality Management

- Database profiling, data auditing
- Integration of data
  - Source selection
  - Data cleaning, ETL
  - Schema and data mapping
  - Record linkage, deduplication
  - Conflict resolution, data fusion
- Constraint and integrity checking
- Data refreshment and synchronization policies
- Metadata management
From the KDD perspective

- Data Quality Mining is beyond data preparation

- Exploratory Data Analysis
- Multivariate Statistics
- Anomaly detection
- Classification
  - Rule-based
  - Model-based
- Clustering
  - Distance-based
  - Density-based
- Visualization
- Quantitative Data Cleaning
  - Distribution transformation
  - Treatment of missing values, inconsistencies, and outliers
What is Data Quality Mining?

“DQM can be defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases.” [Hipp, Güntzer, Grimmer, 2001]

In addition,

Data Quality Mining (DQM) intends to be an iterative framework for creating, adapting, and applying data mining techniques for the discovery, explanation and quantitative cleaning of data glitches and their complex patterns in large and patchy datasets.
Motivation

Data quality problems are:
- Omnipresent in every application domain
- Interwoven and complex in any DB, DW or IS
- Critical to every data management, KDD and decision making project because of their massive financial impact

Limitations of current tools:
- They are *ad-hoc*, specialized, rule-based, and programmatic
- They are specific to a single-type of data quality problem
- They don’t catch interdependences between data quality problems
- Detection and cleaning tools are disconnected
Key Challenges

- Complexity and dimensionality
  - The exact notion of data quality is multidimensional and different from one application domain to another
  - Concomitant data quality problems increase the detection complexity

- Ambiguity
  - The boundary between quality and non-quality data is not precise
  - The boundary between a legitimate anomaly and a data quality problem is hard to define

- Change
  - Data and so data quality keep evolving

- Missing Metadata
Part I. Introduction to Data Quality Research

1. Illustrative Examples
2. Definitions, concepts and motivation
3. Current solutions and their limits
Current Solutions in Practice

- Diagnostic Approaches
  - Database profiling
  - Exploratory data analysis (EDA)

- Corrective Approaches
  - Extract-Load-Transform (ETL)
  - Record linkage (RL)
  - Quantitative Cleaning
Database Profiling

Include descriptive information
- Schema, table, domain, data sources definitions
- Business objects, rules and constraints
- Synonyms and available metadata

Systematically collect summaries of the dataset
- Number of tables, records, attributes
- Number of unique, null, distinct values for each attribute
- Skewness of data distributions
- Field Similarity (Bellman [Dasu et al., 2002])
  - By exact match
  - By substring similarity
    - Resemblance of Q-gram signatures
    - Resemblance of Q-gram min-hash distributions
- Finding Keys and FDs

Mainly applied to relational data
Exploratory Data Analysis (EDA)

**EDA**
- Use of simple statistical techniques for exploring and understanding the data (John Tukey)
- Usually for variable and model selection and for testing distributional assumptions

**EDA for Data Quality**
- Detect data glitches
  - Outliers and extremes
  - Missing values
  - High frequency values and duplicates
- Data transformation for model fitting
- Treatment of glitches
  - Selecting variables and records
  - Replacing using statistical models
EDA – Outlier Detection

- **Control chart/error bounds methods**
  - e.g., expected value; confidence interval or error bounds; 3-Sigma, Hampel bounds, IQR

- **Model-based outlier detection methods**
  - e.g., regression model: outlyingness measured through residuals that capture deviation from the model

- **Multivariate statistics for outlier detection**
  - e.g., density-based and geometric or distance-based outlier detection
EDA - Control chart/error bounds

- Typical value (green) – arithmetic mean, median
- Error bounds (red) – standard deviation, IQR

⚠️ Underlying assumptions of normality and symmetry

⚠️ Simple, but potential for misleading conclusions

⚠️ Non trivial to extend to higher dimensional space
EDA - Model-based outlier detection

- Model captures relationships between variables
- Confidence bounds/bands capture variability
- Points that lie outside bounds

The choice and correctness of the model are critical

Expertise required for parameterization

Finding Multivariate Outliers

**INPUT:** An \(d \times d\) dataset

**OUTPUT:** Candidate Outliers

1. Calculate the mean \(\mu\) and the variance–covariance matrix \(\Sigma\)

2. Let \(C\) be a column vector having length \(d\), the square of the Mahalanobis distance to the mean \(\mu\) is given by:

\[
MD^2 = (x - \mu)' \Sigma^{-1} (x - \mu) = (x - \mu)' \begin{bmatrix} \sigma_{11} & \sigma_{21} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}^{-1} (x - \mu)
\]

3. Find points \(O\) in \(C\) whose value is greater than \(\text{inv} \sqrt{\chi_d^2(0.975)}\)

4. Return \(O\).

Mean and standard deviation are extremely sensitive to outliers (Breakdown point=0%)
Robust estimators

**Minimum Volume Ellipsoid (MVE)** [Rousseeuw & Van Zomeren, 1990]

Let the column vector $C$ with the length $d$ ($d > 2$) be the estimate of location and let the $d$-by-$d$ matrix $M$ be the corresponding measure of scatter. The distance of the point $x_i = (x_{i1}, \ldots, x_{id})$ from $C$ is given by:

$$D_i = \sqrt{(x_i - C)' M^{-1} (x_i - C)}$$

If $D_i > \sqrt{\chi^2_{0.975,d}}$ then $x_i$ is declared an outlier.

$C$ is center of the minimum volume ellipsoid covering (at least) $h$ points of the data set.

**Minimum Covariance Determinant (MCD)** [Rousseeuw & Driessen, 1999]

Given $n$ data points, the MCD is the mean and covariance matrix based on the sample of size $h$ ($h < n$) that minimizes the determinant of the covariance matrix.

Masking the structure of the group of MV outliers (clustered vs scattered)
EDA - Distance-based outliers

Nearest Neighbour-based Approaches

- A point $O$ in a dataset is an $DB(p,d)$-outlier if at least fraction $p$ of the points in the data set lies greater than distance $d$ from the point $O$.
  
  \[ \text{[Knorr, Ng, 1998]} \]

- Outliers are the top $n$ points whose distance to the $k$-th nearest neighbor is greatest.
  
  \[ \text{[Ramaswamy et al., 2000]} \]

Methods fails

- When normal points do not have sufficient number of neighbours
- In high dimensional spaces due to data sparseness
- When datasets have modes with varying density

Computationally expensive
EDA - Density-based outliers

Method
Compute local densities of particular regions and declare data points in low density regions as potential anomalies

Approaches
- Local Outlier Factor (LOF) [Breunig et al., 2000]
- Connectivity Outlier Factor (COF) [Tang et al., 2002]
- Multi-Granularity Deviation Factor [Papadimitriou et al., 2003]

In high dimensional spaces, LOF values will tend to cluster because density is defined in terms of distance.
Clustering for MV Outlier Detection

- AT&T Special service users ~ 1.67M multiple sessions
- Simple k-means clustering based on 7 variables in 2-D projection plot

- In general, computationally expensive and expert parameterization required
- Very sensitive to initial seeds and distance metrics
- Methods fails:
  - When normal points don’t create any clusters
  - In high dimensional spaces, data is sparse and distances become similar

![Graph showing clusters and outliers]

Long hold times, potentially unusual time of day, low volume
Current Practical Solutions

- Diagnostic Approaches
  - Database profiling
  - Exploratory data analysis (EDA)

- Corrective Approaches
  - Extract-Load-Transform (ETL)
  - Record linkage (RL)
  - Quantitative Cleaning
Extract-Transform-Load and Cleaning

Goals

- Format conversion
- Standardization of values with loose or predictable structure
  - e.g., addresses, names, bibliographic entries
- Abbreviation enforcing
- Data consolidation based on dictionaries and constraints

Approaches

- Machine learning and HMM for field and record segmentation [Christen et al., 2002]
- Constraint-based method [Fan et al., 2008]

Performance and scalability issues of most ETL tools
## Academic and Open Source ETL Tools

<table>
<thead>
<tr>
<th>Name</th>
<th>Main characteristics</th>
<th>Data Transformation</th>
<th>Data Cleaning</th>
<th>Duplicate Detection</th>
<th>Data Enrichment</th>
<th>Data Profiling</th>
<th>Data Analysis</th>
</tr>
</thead>
</table>
| Potter’s wheel           | Detection and correction of errors with data transformations: *add, drop, merge, split, divide, select, fold, format*  
Interactivity, inference of the data structure                                                                 | X                   |               |                     |                 |                |               |
| [Raman et al. 2001]      |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Ajax                     | Declarative language based on logical transformation operators: *mapping, view, matching, clustering, merging*  
3 algorithms for record matching                                                                                                                  | X                   | X             | X                   | X               |                |               |
| [Galhardas et al. 2001]  |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Arktos                   | Graphical and declarative (SQL-like and XML-like) facilities for the definition of data transformation and cleaning tasks, optimization, measures of quality factors                                                                 | X                   | X             |                     |                 |                |               |
| [Vassiliadis 2000]       |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Intelliclean             | Detection and correction of anomalies using a set of rules (*duplicate identification, merge, purge, stemming, soundex, stemming, abbreviation*) - Not scalable                                                                 |                     |               |                     |                 |                |               |
| [Low et al. 2001]        |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Bellman                  | Data quality browser collecting database profiling summaries, implementing similarity search, set resemblance, Q-gram sketches for approximate string matching                                                                 |                     | X             | X                   | X               |                |               |
| [Dasu et al., 2002]      |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Febrl                    | Open source in Python, initially dedicated to data standardization and probabilistic record linkage in the biomedical domain, including Q-gram, sorted NN, TF-IDF methods for record linkage and HMM-based standardization | X                   | X             | X                   | X               | X              |               |
| [Christen, 2008]         |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Pentaho-Kettle           | Open source in Java for designing graphically ETL transformations and jobs such as reading, manipulating, and writing data to and from various data sources. Linked to Weka. Easily extensible via Java Plug-ins | X                   | X             | (X)                 | (X)             | (X)            | (X)           |
| [http://kettle.pentaho.org](http://kettle.pentaho.org) |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
| Talend Open Studio       | Open source based on Eclipse RCP including GUI and components for business process modeling, and technical implementations of ETL and data flows mappings. Script are generated in Perl and Java code. | X                   | X             | (X)                 | (X)             | (X)            | (X)           |
| [http://www.talend.com](http://www.talend.com) |                                                                                                                                                                                                                     |                     |               |                     |                 |                |               |
Commercial ETL Tools

Criteria
Ability to execute
- Product/Service
- Overall Viability
- Sales Execution/Pricing
- Market Responsiveness
- Track Record
- Marketing Execution
- Customer Experience
- Operations

Completeness of vision
- Market Understanding
- Marketing Strategy
- Sales Strategy
- Offering (Product) Strategy
- Business Model
- Vertical/Industry Strategy
- Innovation
- Geographic Strategy

Source: Magic Quadrant for Data Integration Tools, Sept. 2008, Gartner RAS Core Research Note G00160825.
Record Linkage (RL) [Elmagarmid et al., 2007]

1. Pre-processing: transformation and standardization
2. Select a blocking method to reduce the search space partitioning the dataset into mutually exclusive blocks to compare
   • Hashing, sorted keys, sorted nearest neighbors
   • (Multiple) Windowing
   • Clustering
3. Select and compute a comparison function measuring the similarity distance between pairs of records
   • Token-based: N-grams comparison, Jaccard, TF-IDF, cosine similarity
   • Edit-based: Jaro distance, Edit distance, Levenshtein, Soundex
   • Domain-dependent: data types, ad-hoc rules, relationship-aware similarity measures
4. Select a decision model to classify pairs of records as matching, non-matching or potentially matching
5. Evaluation of the method (recall, precision, efficiency)
Chaining or Spurious Linkage

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AT&amp;T</td>
<td>180 Park. Av Florham Park</td>
</tr>
<tr>
<td>2</td>
<td>ATT</td>
<td>180 park Ave. Florham Park NJ</td>
</tr>
<tr>
<td>3</td>
<td>AT&amp;T Labs</td>
<td>180 Park Avenue Florham Park</td>
</tr>
<tr>
<td>4</td>
<td>ATT</td>
<td>Park Av. 180 Florham Park</td>
</tr>
<tr>
<td>5</td>
<td>TAT</td>
<td>180 park Av. NY</td>
</tr>
<tr>
<td>6</td>
<td>ATT</td>
<td>180 Park Avenue. NY NY</td>
</tr>
<tr>
<td>7</td>
<td>ATT</td>
<td>Park Avenue, NY No. 180</td>
</tr>
<tr>
<td>8</td>
<td>ATT</td>
<td>180 Park NY NY</td>
</tr>
</tbody>
</table>

Expertise required for method selection and parameterization
## RL - Models and Prototypes

<table>
<thead>
<tr>
<th>Decision Model (Prototype)</th>
<th>Authors</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM-based Method</td>
<td>[Dempster et al. 1977]</td>
<td></td>
</tr>
<tr>
<td>Induction Model</td>
<td>[Bilenko et Mooney 2003]</td>
<td></td>
</tr>
<tr>
<td>Clustering Model (Tailor)</td>
<td>[Elfeky et al. 2002]</td>
<td></td>
</tr>
<tr>
<td>1-1 matching</td>
<td>[Winkler 2004]</td>
<td></td>
</tr>
<tr>
<td>Bridging File</td>
<td>[Winkler 2004]</td>
<td></td>
</tr>
<tr>
<td>Sorted Nearest Neighbors and variants</td>
<td></td>
<td>Empirical</td>
</tr>
<tr>
<td>XML object Matching</td>
<td>[Weiss, Naumann 2004]</td>
<td></td>
</tr>
<tr>
<td>Hierarchical Structure (Delphi)</td>
<td>[Ananthakrishna et al. 2002]</td>
<td></td>
</tr>
<tr>
<td>Matching Prediction based on clues</td>
<td>[Buechi et al. 2003]</td>
<td>Knowledge-based</td>
</tr>
<tr>
<td>Instance-based functional dependencies</td>
<td>[Lim et al. 1993]</td>
<td></td>
</tr>
<tr>
<td>Transformation Functions (Active Atlas)</td>
<td>[Tejada et al. 2001]</td>
<td></td>
</tr>
<tr>
<td>Variant of NN based on rules for identifying and merging duplicates (Intelliclean)</td>
<td>[Low et al. 2001]</td>
<td></td>
</tr>
</tbody>
</table>

Sources:
- Dempster et al. 1977
- Fellegi & Sunter 1969
- Elfeky et al. 2002
- Lim et al. 1993
- Tejada et al. 2001
- Winkler 2004
- Buechi et al. 2003
- Bilenko et Mooney 2003
- Ananthakrishna et al. 2002
- Weiss, Naumann 2004
- Buechi et al. 2003
- Dempster et al. 1977
Interactive Data Cleaning

- **D-Dupe** [Kang et al., 2008] [http://www.cs.umd.edu/projects/linqs/ddupe](http://www.cs.umd.edu/projects/linqs/ddupe)
  Duplicate search and visualization of cluster-wise relational context for entity resolution

- **Febrl** [Christen, 2008]: [https://sourceforge.net/projects/febrl/](https://sourceforge.net/projects/febrl/)
  Rule-based and HMM-based standardization and classification-based record linkage techniques

- **SEMANDAQ** [Fan et al., 2008]: CFD-based cleaning and exploration

- **HumMer** [Bilke et al., 2005]: Data fusion with various conflict resolution strategies

- **XClean** [Weis, Manolescu, 2007]: Declarative XML cleaning
Commercial Data Quality Tools

DQ Tools include:
• Profiling
• Improvement:
  Standardization
  Cleansing
  Matching
  Enrichment
• Monitoring

Gartner RAS Core Research Note G00157464
Quantitative Data Cleaning

**Methods**

- **Inclusion** *(applicable for less than 15%)*
  - Anomalies are treated as a specific category

- **Deletion** *(for less than 2%)*
  - List-wise deletion omits the complete record
  - Pair-wise deletion excludes only the anomaly value from a calculation

- **Substitution** *(applicable for less than 15%)*
  - Single imputation based on mean, mode or median replacement
  - Linear regression imputation
  - Multiple imputation (MI)
  - Full Information Maximum Likelihood (FIML)
Limits of Current Methods

Non Realistic Assumptions

- Data quality problems may not be rare events → rare class mining won’t give the “complete picture”
- Data quality problems don’t occur at random → MCAR/MAR assumptions are not applicable
- Data quality problems are not uniformly distributed → model-based assumptions is hazardous
- Different types of data quality problems may co-occur and be (partially) correlated → mutual masking-effect of concomitant DQ problems → potential multicollinearity problem
- Their (co-)occurrence should be explainable → explanatory variables/processes may be external and out of the scope of the analysis
- They should be corrected with “predictable side-effects” → Biases of imputation and regression methods
In Particular

**EDA - Statistics**
- Methods tied to known distributions
- Parametric assumptions often do not hold for real datasets
- Bad points can completely skew the mean and standard deviation: Robustification of the estimators is required
- Statistical methods suffer the uni-modality and locality assumptions: they consider the data set globally

**EDA - Clustering**
- Magic numbers,
- Complex parameterization and settings
- Locality and normality assumptions

**Quantitative Cleaning**
- Only applicable when the dataset has less than 15% of glitches
- Non negligible biases
Challenges of DQM

Detection of concomitant DQ problems

- Joint detection of fuzzy duplicates, disguised missing values, multivariate outliers, and deviants
- Detection of complex patterns of multivariate glitches
- Interactivity, rerunability and recoverability of DQM processes

DQM in high dimensional data sets

Benchmarking
Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study
Part II. Data Quality Mining

1. DQM Framework
2. Advances on Single-Type DQ Problems
3. Discovery of Complex Glitches
Data Quality Mining: The Big Picture

Data Quality Management

Data Quality Mining

Data Source Selection
Data Quality Requirements

Source DB Profiling

Detection & Handling DQ Problems

Extract Transform Load Data

Recommendations and Corrective Actions

Data Staging Area

Upstream the KDD Process: Warehousing

Downstream the KDD Process: Decisional Mining

Objectives:
Data Quality Measurement and Improvement
DQM Framework: The Process

Iterative Detection and Cleaning

- Inconsistent Data
  - Constraint
- Missing Data
  - Imputation
- Duplicates
  - De-duplication
- Outliers
  - Uni and MV Detection

Patterns and Dependencies among Glitches

Best DQM Strategy?
Best DQM Strategy

From your patchy raw data, Define:

- An ideal dataset
  - Baseline; historical aggregate; pre-defined
- The representation
  - Signature; a statistical summary
- The comparison measure
- The function to optimize
  - Min, threshold

Plan the cleaning process to get the optimal dataset
Part II. Data Quality Mining

1. DQM Framework
2. Advances on Single-Type DQ Problems
   - Inconsistent Data
   - Deduplication
   - Missing Data
   - Outlier Mining
3. Discovery of Complex Glitches
Inconsistent Data

- **Probabilistic Approximate Constraints** [Korn et al., 2003]
  
  Given a legal ordered domain on an attribute,
  
  - A **domain PAC** specifies that all attribute values \( x \) fall within \( \varepsilon \) of \( D \) with at least probability \( \delta \), as \( \Pr(x \in [D \pm \varepsilon]) \geq \delta \)
  
  - A **functional dependency PAC** \( X \rightarrow Y \) specifies that, if
    
    \[
    |T_i.A_\ell - T_j.A_\ell| \leq \Delta_\ell \quad \forall A_\ell \in X \text{ then } \Pr\left(|T_i.B_\ell - T_j.B_\ell| \leq \varepsilon_\ell \right) \geq \delta \quad \forall B_\ell \in Y
    \]

- **Pseudo-constraints** [Ceri et al., 2007]
  
  Pair \(<P1,P2>\) where \( P1 \) and \( P2 \) are predicates on the same domain \( D \) such that if \( P1 \) holds, then usually \( P2 \) also and therefore there are few rule violations. More formally, based on the probability contingency table,
  
  \[
  \frac{p_{11}}{p_{11} + p_{21}} - \rho - (1 - \rho) \cdot (p_{11} + p_{12}) > 0
  \]

- **Pattern Tableaux for Conditional Functional Dependencies**
  

  A CFD is defined to be a pair \( \varphi = R(A \rightarrow B, T_\rho) \), where \( T_\rho = \)

  \[
  \begin{array}{c|c|c}
  A & B & \text{Embedded FD} \\
  \hline
  - & b_1 & \phantom{0} \\
  - & b_2 & \phantom{0}
  \end{array}
  \]
# Duplicate Data: Learning Approaches for RL

## Training examples

<table>
<thead>
<tr>
<th>Customer</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Similarity distance functions

<table>
<thead>
<tr>
<th>$f_1$</th>
<th>$f_2$</th>
<th>...</th>
<th>$f_n$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.4</td>
<td>...</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>...</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
<td>...</td>
<td>0.4</td>
<td>1</td>
</tr>
</tbody>
</table>

## Unlabeled list

<table>
<thead>
<tr>
<th>Customer</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Learnt Rule:

\[
\text{Learnt Rule: All-Ngrams} \times 0.4 \\
+ \frac{\text{CustomerAddressNgrams}}{0.2} \\
- 0.3 \times \text{EnrollYearDifference} \\
+ 1.0 \times \text{CustomerNameEditDist} \\
+ 0.2 \times \text{NumberOfAccountsMatch} - 3 > 0
\]

### Learners:

- SVMs: high accuracy with limited data [Christen, 2008]
- Decision trees: interpretable, efficient to apply
- Perceptrons: efficient incremental training [Bilenko et al., 2005]
Assumption: On an attribute, there often exist only a small number of inliers whose values are frequently used as disguised missing data, e.g., popular DoB: Jan 1.

For a tuple $t$ in the truth table $T$ and $\tilde{t}$ in the recorded table $\tilde{T}$, the value $\tilde{t}.A$ is a disguised missing value if $t.A = \otimes$ but $\tilde{t}.A \neq \otimes$.

Embedded Unbiased Sample Heuristic

Goal: Find the largest set $M_v$ embedding $S_v$, and maximizing a correlation-based sample quality score.
Handling Missing Data

- **Completion Using Association Rules**
  - Based on a consensus from rules with high confidence and user interaction
  - Based on measures scoring the best rules to select the replacement value [Wu et al., 2004]

- **Imputation using NN, Clustering and SVM**
  - K-Nearest Neighbour Imputation [Batista, Monard, 2003]
  - K-means Clustering Imputation [Li et al., 2004]
  - Fuzzy K-means Clustering [Acuna, Rodriguez, 2004]
  - SVM [Feng et al. 2005]
Outlier Mining

- Multivariate techniques
  - Projection pursuit
  - Distance and depth based methods
  - Probability and kernel based methods
- Stream specific methods
- Too many outliers $\rightarrow$ Distributional shift?
  - Change detection
- Great tutorial on outliers [Kriegel et al., 2009]:
  http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/tutorial_slides.pdf
Projection Based Methods

- Projection pursuit techniques are *applicable in diverse data situations* although at the expense of high computational cost.
  - No distributional assumptions, search for useful projections
- **Time Series**: Galeano et al. (2006) extend the idea of projecting in directions of high and low kurtosis to multivariate time series.
- **Skewed Distributions**: Hubert and Van der Veeken (2007) extend the boxplot idea by defining adjusted outlyingness followed by random projections for detecting outliers in skewed data.
**Outlier Mining - Robust PCA**

[Shyu et al., 2003]

**INPUT:** An $N \times d$ dataset

**OUTPUT:** Candidate Outliers

1. Compute the principal components of the dataset
2. For each test point, compute its projection on these components
3. If $y_i$ denotes the $i^{th}$ component, then the following has a chi-square distribution

$$
\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} + \cdots + \frac{y_q^2}{\lambda_q}, q \leq p
$$

3. For a given significance level $\alpha$, an observation is an outlier if

$$
\sum_{i=1}^{q} \frac{y_i^2}{\lambda_i} \geq \chi_q^2(\alpha)
$$
Outlier Identification in High Dimensions
[Filzmoser, Maronna and Werner, 2008]

- Works in very high-D, where dimensions > samples, e.g., gene data
- Differential weights to detect location and scatter outliers; weights combined in final step
- Based on robust statistics
Outlier Detection for Skewed Data

[Hubert and Van der Veeken, 2007]

- For skewed distributions
- Key concepts
  - Adjusted outlyingness – different scaling on either side of median in boxplots.
  - MV equivalent, e.g., bagplot in 2-D
  - Random projections to identify outliers
Distance and Depth Based Methods

- Distance-based methods aim to detect outliers by computing a measure of how far a particular point is from most of the data.

- *Robust methods*
  - Robust distance estimation in high-D [Maronna and Zamar, 2002] [Pena and Prieto, 2001]

- *Depth based nonparametric methods*
  - Nonparametric methods based on multivariate control charts [Liu et al, 2004]
  - Outlier detection with kernelized spatial depth function [Cheng, Dang, Peng and Bart, 2008]

- *Exotic methods*
  - Angle based detection [Kriegel, 2009]
**DDMA: Nonparametric Multivariate Moving Average Control Charts Based on Data Depth**

[Liu, Singh and Teng, 2004]

- Extends simplicity of control charts to higher dimensions – relatively few assumptions
- Use any data depth, e.g., simplicial depth to map multidimensional data to a scalar and rank
- Apply moving average control chart techniques to data depth rank to identify outliers

Deepest point, e.g., simplicial depth = contained in most triangles
Probability and Kernel based methods

- **Popular methods**: LOF, INFLO, LOCI  
  see Tutorial of [Kriegel et al., 2009]
- **Mixture distribution**: Anomaly detection over noisy data using learned probability distributions [Eskin, 2000]
- **Entropy**: Discovering cluster-based local outliers [He, 2003]
- **Projection into higher dimensional space**: Kernel methods for pattern analysis [Shawne-Taylor, Cristiani, 2005]
Probability Based Methods

- **Probability distributions** [Eskin, 2000]
  
  **Assumption:** High probability to have the number of normal elements in a dataset $D$ significantly larger than the number of outliers

  **Approach:**
  From the distribution for the dataset $D$ given by: $D = (1-\lambda) M + \lambda A$
  with $M$: Majority distribution and $\lambda$: Anomaly distribution
  - Compute likelihood of $D$ at time $t$: $L_t(D)$
  - Measure how likely each point $p_t$ is outlier at time $t$ such as:
    $$M_t = M_{t-1} \setminus \{p_t\} \text{ and } A_t = A_{t-1} \cup \{p_t\}$$

- **Entropy-based methods** [Lee et al., 2001][He 2005]
  
  **Observations:** large entropy $\rightarrow$ partition into regular subsets
  skewness of class distribution $\rightarrow$ small entropy $\rightarrow$ high redundancies

  **Approach:**
  Find a k-sized subset whose removal leads to the maximal decreasing of entropy
Kernel methods for pattern analysis

[Shawne-Taylor, Cristani, 2005]

Objective: Mapping data by a nonlinear mapping function $\Phi$ into a higher feature space where they are linearly separable into majority and outliers.
Stream Specific Methods

- **Distance based outliers**: Detecting distance based outliers in streams of data. [Anguilli and Fassetti, 2007]
- **Distributed streams**: Adaptive Outlier Detection in Distributed Streams [Su, Han, Yang, Zou, Jia, 2007]
- **A general density estimation scheme**: Online outlier detection in sensor streams [Subramaniam et al., 2006]
- **Projections and high dimensions**: Projected outliers in High-D data streams [Zhang, Gao, Wang, 2008]
- **Items of interest**: Finding frequent items in data streams [Cormode and Hadjieleftheriou, 2008]
Online Outlier Detection in Sensor Data Using Non-Parametric Models
[Subramaniam et al., 2006]

- Online outlier detection in hierarchical sensor networks
- Solve the more general problem of estimating the multidimensional data distribution
  - Chain sampling
  - Epanechnikov kernel
Outliers and Change Detection

- Often, an increase or decrease in outliers is the first sign of a distributional shift
- Serious implications for data quality – recalibrate anomaly detection methods
- Change detection methods are critical
Change Detection Schemes

- **Comprehensive framework:** Detecting Changes in Data Streams. [Kifer et al., 2004]
- **Kernel based:** Statistical Change Detection in Multi-dimensional Data. [Song et al., 2007]
- **Nonparametric, fast, high-D:** Change Detection you can believe in: Finding Distributional Shifts in Data Streams. [Dasu et al., 2006, 2009]
Change (Detection) you can believe in:
Finding Distributional Shifts in Data Streams
[Dasu, Krishnan, Li, Venkatasubramanian, Yi, 2009]

- Compare data distributions in two windows
  - Kdq-tree partitioning
  - Kullback-Leibler distance of histograms
    - Counts
    - Referential distance
  - Bootstrap to determine threshold
- File descriptor data stream
  - 3 variables shown
  - Change detection led to improvement in process and cycle times
Part II. Data Quality Mining

1. DQM Framework
2. Advances on Single-Type DQ Problems
3. Discovery of Complex Glitches
Changes in Distributions Caused by Missing/Duplicate Data

- Subtle cases of duplication/missing data
  - Result in changes in distributions
  - Missing $\rightarrow$ “lower” density regions
  - Duplicates $\rightarrow$ “higher” density regions

- Multinomial tests
  - Contingency tables (Chi-square test)
  - Difference in proportions (e.g., counts)

- Difference in Distributions
  - Histogram distances (Kullback Leibler)
  - Rank based (Wilcoxon)
  - Cumulative distribution based (Kolmogorov-Smirnov)
Missing Data Example

- Comparison of telecommunications data sets
- Anomalous months
  - Missing data
  - Kdq tree partition
  - Darker $\rightarrow$ greater density difference
- Automatic detection is speedy, provides an opportunity to recover and replace data before it is archived
Outline

Part I. Introduction to Data Quality Research

Part II. Data Quality Mining

Part III. Case Study
Case Study: DQ Patterns in networking data streams

- Domain knowledge and goal
- Data set description
- DQM tasks
- Analysis
- Best DQM strategy
IP Network Data Streams

- Thousands of network elements on an IP network
- Transmit data streams as they communicate to verify availability and transmit data
- Monitor the data streams to measure performance, and to optimize the network
IP Data Streams: Attributes

- Data measured at desired frequency
  - Real-time, minutes, hours
- Massive amounts of data!
- Attributes
  - Resource usage
  - Traffic measurements
  - Performance metrics
  - Alarms
- Gathered from multiple, disparate sources
IP Data Streams: A Picture

- 10 Attributes, every 5 minutes, over four weeks
- Axes transformed for plotting
- Glitches
  - Automatic detection
IP Data: Multivariate Glitches

Outliers

Missing

Duplicates
DQM Tasks: IP data streams

- **Goal**
  - Extract the “cleanest” dataset
  - Discover patterns and interactions between glitches

- **Challenges**
  - Domain specific: dynamic nature of the network
  - Data integration

- **Focus**
  - Missing Data
  - Duplicates
  - Outliers
  - Others are not considered in the case study
IP Data: Data Quality Mining

Input Data Stream

Dataset Profiling
EDA
Visualization

Clean Data Stream

Mining/Cleaning Strategies

Missing Data
Imputation
Point Estimate/Regression/Joint Density

Duplicates
De-duplication
Unique Key/Unique Record
Random, Most Recent/Averaging

Patterns
Lag variables/Association Rule Discovery

Best DQM Strategy?
Best Strategy Definition

- **Ideal data**: Super clean data, any record that has no ambiguity or imperfection
- **Representation**: A histogram (MV histogram using kdq-tree)
- **Distance**: Kullback Leibler distance
- **Best strategy**: Smallest KL distance to the Ideal
Finding the optimal dataset

MV strategy shown with Univariate projection

0 = Raw data
1 = Median-based imputation from 4
2 = De-duplicated data
3 = 1 + 2
4 = Super clean (Full deletion)
5 = Strategy driven by the discovery of patterns of glitches
What are the patterns?

Types of Patterns

- Univariate/Multivariate
- Discovered Patterns
  - Missing-duplicate pairs: the responses arrive at the bin boundary resulting in the pattern
  - Complex patterns:
    - co-occurrence or lagged occurrence of outliers and missing,
    - outliers and duplicates
    - missing and duplicates
- Cleaning strategies
  - Quantitative cleaning, e.g., blind imputation
  - Domain knowledge-driven replacement of missing values with adjacent duplicates
  - Additional iterations needed because cleaning reveals new glitches
Case Study: Conclusion

- IP data stream – multivariate, massive, glitchy
- Critical for network monitoring
- Patterns and dependencies in glitches are used to recover much of the data such that the treated dataset is close to the ideal dataset
- Discovery of explanatory variables is useful for understanding recurrent DQ problems
In Summary
DQM Summary: Multivariate Glitches

- Glitches are multivariate with strong interdependencies
  - Static & temporal
  - Domain and application dependent
- DQM framework is important
  - Extant approaches tend to treat each class of glitches separately – misleading.
- Patterns and distribution of glitches are crucial in formulating cleaning strategies
DQM Summary: Process and Strategies

- Iterative and complementary cleaning strategies
- Best DQM strategies
  - Quantitative criteria
  - Resource-dependent
  - Domain, user and operational needs

Iterative Detection and Cleaning

- Inconsistent Data Constraint
- Missing Data Imputation
- Duplicates Deduplication
- Outliers Uni- and MV- Detection

Patterns and Dependencies among Glitches
Thanks

Any questions?
Data Quality: 
Up-coming Events

August 24, 2009:  
**QDB (Quality in Databases)**  
Workshop in conjunction with VLDB 2009 in Lyon, France  
[http://qdb09.irisa.fr](http://qdb09.irisa.fr)

November 7-9, 2009:  
**ICIQ (International Conference on Information Quality)**  
Hasso-Plattner Institut, Potsdam, Germany  
[http://www.hpi.uni-potsdam.de/naumann/iciq2009/](http://www.hpi.uni-potsdam.de/naumann/iciq2009/)
Limited Bibliography
References

Books
- Tukey, John Wilder. Exploratory Data Analysis. Addison-Wesley, 1977

Surveys
References

Tutorials


Data Profiling


Data Preparation and Data Quality Mining

References

Data Cleaning – ETL

- HERNANDEZ, M., STOLFO, S., Real-World Data is Dirty Data Cleansing and the Merge/Purge Problem, Data Mining and Knowledge Discovery, 2(1)9-37, 1998.
References

Record Linkage and duplicate detection (1/2)

References

Record Linkage and duplicate detection (2/2)

References

Inconsistencies

- FAN, Wenfei, GEERTS, Floris, JIA, Xibei Semandaq A Data Quality System Based on Conditional Functional Dependencies, VLDB'08, (demo), 2008.

Change Detection

Outlier Detection (1/2)

Outlier Detection (2/2)

- Su, L., Han, W., Yang, S., Zou, P., Jia, Y., Continuous adaptive outlier detection on distributed data streams. In HPCC, LNCS 4782, pp. 74-85, 2007.
References

Missing Values